Université de Toulouse — École Doctorale Informatique et Télécommunications

Formation Doctorale Signal, Image, Acoustique, Optimisation

Caractérisation de l'environnement musical dans les documents audiovisuels

THÈSE

présentée et soutenue publiquement le 8 Décembre 2009

pour l'obtention du

Doctorat de l'Université de Toulouse

(spécialité Informatique)

par

Hélène LACHAMBRE

Composition du jury

Rapporteurs :	Mme Annie MORIN	IRISA – Université de Rennes 1
	M. Gaël RICHARD	ENST – TELECOM ParisTech
Examinateur :	M. Geoffroy Peeters	IRCAM
Directrice de thèse :	Mme Régine ANDRÉ-OBRECHT	IRIT – Université de Toulouse
Encadrant :	M. Julien PINQUIER	IRIT – Université de Toulouse
Présidente :	Mme Corinne Mailhes	IRIT – INP Toulouse

Institut de Recherche en Informatique de Toulouse — UMR 5505

Remerciements

Mardi 15 décembre, j'ai soutenu ma thèse depuis exactement une semaine. Autant dire que le plus dur est fait. Est-ce bien sûr ? La tâche à laquelle je m'attelle maintenant – l'écriture des remerciements – est peut-être la plus périlleuse de ces trois années : « surtout n'oublier personne, avoir un mot pour chacun, écrire quelque chose d'original... ».Tout cela en tenant compte du fait que pour de nombreuses personnes que je vais citer ici, ces deux pages seront peut-être les seules qu'elles liront jamais de cet ouvrage !!!

Je prie donc d'avance toutes les personnes que j'aurais pu oublier de bien vouloir me pardonner, et si elles le veulent, de rajouter – à la main – leur nom dans le paragraphe approprié dans leur exemplaire de cet ouvrage.

Je commencerai ces remerciements par Régine, directrice de thèse formidable, à la fois pour m'avoir proposé ce sujet – et avant un stage de master 2 –, mais surtout pour la qualité de son encadrement tout au long de ces trois années. Elle m'a apporté certes des idées, mais aussi des connaissances plus générales sur le milieu de la recherche et de l'enseignement supérieur, son fonctionnement...

Juste après vient Julien, qui a également participé à l'encadrement de ma thèse. Ses conseils, idées, mais aussi son aide pour tous les détails administratifs de la vie d'un chercheur, m'ont vraiment facilité la vie. Et tes blagues récurrentes sur les frites et les litchis n'ont rien enlevé, bien au contraire, au plaisir que j'ai eu à travailler avec toi.

Viennent maintenant M. Gaël Richard et Mme Annie Morin, qui ont accepté d'être rapporteurs sur ma thèse. Merci pour vos commentaires, et pour avoir participé au jury de ma soutenance. Dans mon jury, j'ai également eu le plaisir d'avoir M. Geoffroy Peeters, dont les questions m'ont ouvert de très nombreuses pistes de travail. Enfin, merci beaucoup à toi, Corinne, d'avoir accepté de présider ce jury.

Pour la réussite d'une thèse, il faut certes compter sur un bon sujet et un bon encadrement, mais l'environnement de travail compte également énormément. Sur ce point, il me semble être plutôt bien tombée. Je salue toute l'équipe SAMoVA, soit Benjamin, José et Élie, qui m'ont supporté comme collègue de bureau, mais aussi (dans l'ordre, en partant du bout du couloir) Philippe (le chef), Thomas, Isabelle, Brice, Philippe (le petit dernier), Lionel, Ioannis, Reda, Jérémy (« Mr. Performances Alternatives », on se comprend...), Fred, Khalid, Jérôme (qui a si souvent, avec Julien, dépanné mon ordi), Jérôme, Christine, et Hervé. À force de volonté, vous avez même fini par faire de moi une informaticienne !

Ici se place un grand merci à M. Didier Dubois qui, m'a ouvert sa bibliothèque musicale. C'est de là que provient la grande majorité des extraits de chant *a capella* utilisés ici.

J'étais Monitrice, je suis désormais ATER, tout ceci à l'N7. Je ne saurais donc passer sous silence la qualité d'enseignement de toute l'équipe « Traitement du Signal », que j'avais pu constater en tant qu'étudiante, et que j'ai confirmé en tant que collègue. Je salue en particulier les collègues avec lesquels j'ai partagé des cours : JR (binôme de choc !), Olivier, Jérôme (encore un ?), Marie, Corinne (encore elle !), Richard (notre maître à tous en DSP), Benjamin (le même que plus haut), Reza, Nathalie, Caroline, Annie, et PA. Merci également à Danielle, pour m'avoir acceptée sur ces deux postes, et à Cathy, la secrétaire rêvée du département électronique.

Enfin, au cours du monitorat, j'ai eu l'occasion – comme tous les moniteurs – de faire un « atelier projet ». Ce fut l'occasion de rencontrer quatre personne, qui sont maintenant devenues des amis : Solenn (alias Fisie, la chimiste), Aurélie (la physicochimiste), Sébastien (alias Le Grand Sorcier, le physicien), et Jessica (la légiste).

Passons maintenant aux personnes qui ont contribué de façon plus indirecte au succès de ces trois années.

Les amis, bien sûr.

Les grimpeurs, pour les sorties le week-end, les week-end prolongés, mais aussi les soirées sans grimpe : Tom, Cissou, Ben, Caro, Alice, Laurent, Laurent, et Clément.

Les joueurs, qui parfois émargent également dans le paragraphe précédent : Manu, Domi, Pierre, Pauline, Ben, Caro, Tom, et Cissou.

Les montagnards de tout poil, skieurs, marcheurs,. . . , souvent anciens de l'N7, et qui, là encore, appartiennent déjà pour certains à l'une, l'autre ou les deux des précédentes catégories : Cat, François, Clément, Laure, Laurène, Greg, Ben, Caro, Manu, et Domi. Les nouvel an prolongés au ski, les grands week-end ensemble, c'est quand vous voulez !

Les inclassables : Sandrine, Nico, Zitoun, Francis (celui qui le premier m'a fait aimer la musique), JR, et Steph.

Les deux indispensables : Maud (et Pierre), et Claire (et Éloi, et depuis peu Zélie), un an en colloc' en prépa, ça ne s'oublie pas !

La famille, évidemment.

Un grand merci à ma Maman pour sa relecture de ma thèse, pour toutes les erreurs de frappe, d'orthographe et de grammaire restantes qu'elle a corrigées et pour toutes les remarques pertinentes qu'elle m'a faites.

Encore merci à ma Maman, et à mon Papa, d'avoir traversé toute la France juste pour assister à ma soutenance. Quels parents fabuleux vous avez été jusqu'ici, et vous serez j'en suis sûre ensuite. J'ai la chance de vous avoir !

Olivier, j'en suis sûre, fera partie des personnes qui dépasseront les remerciements et liront cet ouvrage jusqu'au bout. J'en profite pour te remercier encore une fois de ta relecture attentive de mes articles en anglais, tu es mon relecteur « fluent english speaker ». Adeline se lassera peut-être avant la fin de cet ouvrage, mais je sais que je pourrai toujours compter sur son soutient et son amitié sans failles, et que nous continuerons à voyager entre Toulouse et Limoges pour des rencontres fraternelles. Vous êtes tous les deux géniaux comme frère et sœur !

Et pour finir, « last but sûrement not least »...

Merci enfin à toi, Adrien, pour être là tous les jours. Tu y as cru au moins autant que moi (sinon plus !), tu y as cru même quand ce n'était, par moments, plus mon cas, et tu y as cru jusqu'au bout. Ta contribution à cette thèse est difficilement mesurable, mais elle fût absolument indispensable.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 L'environnement de recherche

Depuis plusieurs dizaines d'années, l'indexation *par le contenu* des documents audiovisuels fait l'objet de travaux de la part de nombreuses équipes de recherche :

- Le mot « audiovisuel » fait référence à un contenu « multimédia » qui regroupe à la fois des données audio, des données images ou séquence d'images, des données textuelles...
- Le mot « document » audiovisuel laisse place à une multitude d'objets, même en se limitant aux seuls documents numériques : pages Web, émissions de télévision et de radio, vidéos personnelles en sont autant d'exemples.

Compte tenu de l'étendue de ce domaine de recherche, il était naturel d'aborder le problème de l'indexation par le contenu sous plusieurs angles. Les connaissances pluridisciplinaires ont amené les équipes de recherche à l'appréhender avec les yeux de leur discipline initiale : la communauté image (resp. parole, vidéo, texte) a étudié l'indexation par le contenu des documents image (resp. parole, vidéo, texte). À la fin des années 1990, est apparue une nouvelle piste de recherche, suite à une prise de conscience, à savoir qu'un document audiovisuel est par essence « multimédia » et que les contenus de ces médias sont inévitablement corrélés. C'est ainsi que sont apparus des congrès sur l'indexation par le contenu[1], et des équipes multimédia[2].

L'équipe SAMoVA[3], basée à l'IRIT[4] est née en 2002 de la rencontre d'acteurs issus du monde de la parole et de la vidéo, dans le but de travailler conjointement sur l'indexation par le contenu multimédia de documents. Les travaux réalisés dans l'équipe visent plus précisémment à exploiter la dimension temporelle des deux médias que sont l'audio et la vidéo, et la corrélation entre ces médias.

[1] Le premier congrès CBMI (Content Based Multimedia Indexing) date de 1999
[2] Par exemple l'équipe TEXMEX à l'IRISA (Rennes)
[3] Structuration, Analyse, MOdélisation de documents Vidéo et Audio
[4] Institut de Recherche en Informatique de Toulouse

Nombre de travaux de l'équipe visent à décrire le flux audio au travers de macro segments (parole, musique, locuteur...). Ces travaux se démarquent de la tendance généralement rencontrée en audio, qui consiste à transcrire finement l'audio (transcription de la parole pour atteindre le message prononcé, transcription de la musique pour atteindre la partition). Les résultats de l'équipe SAMoVA en audio ont eu pour objet la détection de la parole et de la musique [PRAO03], la reconnaissance de la langue [RFPAO05], la segmentation et le regroupement en locuteur [EKSP09], la vérification du locuteur [LDB07]. Au niveau applicatif, il s'agit de décrire le type d'information présente, sans en préciser le message exact. Par exemple, on se demande « Qui parle ? », ou « Dans quelle langue ? », sans s'intéresser au message véhiculé. L'objectif ultime est de structurer un flux audiovisuel (flux télévisuel ou radiophonique) en émissions (journal, film, émission culturelle...), et de décrire le contenu de chaque émission.

Au cours de cette démarche, il est apparu que la musique était au sein de l'équipe le « parent pauvre », même si le détecteur de musique proposé lors de la campagne d'évaluation ESTER [GGM+05] a donné de très bons résultats.

Lors de l'analyse du flux audio (télévisuel ou radio), la musique se retrouve dans la composante « non parole », sans aucune distinction de contenu. C'est ainsi que sont regroupés les jingles, les extraits musicaux publicitaires, les génériques de film, les pauses musicales des émissions.

Les travaux de cette thèse ont eu pour but initialement d'explorer cette composante musique et d'essayer d'en extraire des macros segments. La tâche étant complexe, nous nous sommes limités à l'identification des segments monophoniques / polyphoniques, et à l'identification des segments contenant du chant. Notons que ce dernier représente une cause non négligeable d'erreurs lors d'une segmentation parole / musique. Il serait alors intéressant de le considérer comme une classe à part entière, pour améliorer la détection des composantes primaires (parole, musique, chant, bruit...).

1.2 L'indexation de la musique et par la musique

L'indexation de la musique est un des défis majeurs actuels. En effet, des millards d'heures de musique sont produites chaque année, dont une grande partie est d'ailleurs mise en ligne. De nombreux enjeux sont liés à cette diffusion :

- Des enjeux pour les professionnels, avec la détection de copies, ou la promotion d'artistes via les sites sociaux[5].
- Des enjeux pour les particuliers, avec les problèmes liés à la recherche de titres de musique.

En effet, si l'utilisateur cherche en particulier un morceau, dont il connaît le titre, ou encore la musique d'un compositeur dont il connaît le nom, l'indexation actuelle par les

[5]Par exemple le site http://www.myspace.com/

méta-données associées aux fichiers musicaux est suffisante. Ces méta-données (titre de la chanson ou de l'album, noms des compositeur, interprète, maison de disque...) sont au moins en partie renseignées pour une grande part de la musique numérisée.

En revanche, pour des recherches plus vagues, une analyse informatique de la musique elle-même est nécessaire. Même pour des requêtes relativement ciblées, telles que « je cherche un morceau de J.-S. Bach, en Do mineur », la seule utilisation de la méta-donnée *compositeur : J.-S. Bach* n'est pas forcément suffisante, ce compositeur ayant été particulièrement prolixe. Que dire alors de la difficulté pour une recherche du type « je cherche une musique Rock joyeuse, en Ré Majeur, avec un rythme rapide » !

Pour faire face à cette demande pressante, de très nombreux travaux sont actuellement réalisés pour la description automatique de la musique. Ces travaux portent sur des sujets aussi variés que l'identification du style, des émotions ou du chanteur, la détermination de la tonalité ou de la pulsation, ou encore diverses transcriptions de la musique (mélodies, paroles, accords, partitions).

L'indexation par la musique vise à spécifier le rôle de la musique. Elle a une grande importance dans le contenu des documents audiovisuels, et dans la perception que l'auditeur / spectateur en a ; tous les réalisateurs d'émission radio, tous les réalisateurs de films en sont conscients lorsqu'ils choisissent une musique particulière pour une pause musicale à la radio ou un film. On imaginerait par exemple mal le thème musical du film « Les Dents de la Mer » sur les scènes du « Fabuleux Destin d'Amélie Poulain » ! Les informations extraites de la musique proviennent de la musique de fond, des jingles, d'extraits musicaux, et de morceaux de musique complets. Les outils développés pour l'indexation de la musique et par la musique ont une interaction très importante, sans se recouvrir totalement, compte tenu de l'objectif final :

- La recherche d'un jingle n'a aucun intérêt d'un point de vue musicologique.
- La caractérisation d'une émission de variété au travers d'une alternance de chant, parole, musique (sans chant) ne nécessite pas une transcription fine ni de la musique, ni de la voix chantée, ni de la parole.

Les travaux que nous présentons se situent à la frontière de ces deux approches, puisqu'il s'agit de trouver une macro segmentation du flux musical qui affine la détection des morceaux de musique déjà mis en évidence par les segmentations parole / non parole et musique / non musique [PRAO03].

1.3 Positionnement de l'étude

Comme dit précédemment, l'indexation du flux audio, notamment des flux télévisuel et radiophonique, impose de décomposer le signal acoustique en ses composantes primaires, afin d'utiliser ensuite les outils adéquats à chaque type de segment et en extraire un contenu spécifique.

Dans ce type d'application, la détection du chant ou d'une voix chantée s'avère également nécessaire.

La détection du chant est évidemment un pré-traitement nécessaire pour *l'identification du chanteur*, tout comme pour la *transcription des paroles*. Elle peut également servir de pré-traitement dans plusieurs autres tâches.

Lorsqu'il y a du chant, celui-ci tient très souvent la mélodie principale. Il peut être intéressant de disposer de cette information pour adapter les outils de *transcription de la mélodie principale* à ce cas particulier – notons que cela est déjà le cas dans certaines méthodes de transcription.

Enfin, la présence de chant – ou son absence – peut être caractéristique de certains *genres musicaux*. On pensera par exemple au rap ou à l'opéra, dans lesquels la voix est toujours présente, ou, à l'inverse, à de nombreux genres de musique classiques (symphonie, certaines danses comme la valse), dans lesquels il n'y a jamais de chant.

Nos premières études sur le chant ont montré un intérêt à distinguer le chant *a capella* du chant accompagné. Cette remarque nous a conduit à examiner plus généralement la distinction du contexte monophonique et du contexte polyphonique, distinction qui devient dès lors un pré-traitement à la recherche du chant.

Les deux contributions majeures des recherches présentées dans ce document sont :
- la distinction entre sons monophoniques et sons polyphoniques,
- la détection du chant.

1.4 Organisation du mémoire

Afin de s'imprégner de la recherche actuelle en musique, nous commençons ce mémoire par la présentation, dans le chapitre 2, de l'état des recherches pour la description automatique de la musique. Nous nous attardons en particulier sur les outils développés pour en décrire divers aspects : instruments présents, tonalité, rythme, genre, émotions, identité du chanteur, transcriptions de la mélodie, de la partition, des accords, et des paroles.

Les chapitres 3 et 4, sont consacrés aux deux points que nous avons étudiés : d'une part la distinction entre les sons monophoniques et les sons polyphoniques, d'autre part la détection du chant. Pour chacun de ces deux thèmes, nous présentons, au sein de chaque chapitre, l'état de l'art, notre approche, et l'ensemble des expériences que nous avons menées afin de valider celle-ci.

Enfin, dans le chapitre 5, nous revenons sur le potentiel de l'indexation par la musique, et nous présentons quelques réflexions sur l'aide que la musique peut apporter dans l'indexation et la description automatique des flux audio.

Chapitre 2

L'environnement musical - Introduction

Sommaire

2.1 Introduction - environnement sonore, environnement musical

L'« environnement sonore » est un terme mal défini – ou plutôt trop défini. Parmi les définitions proposées dans le domaine du traitement automatique des sons, nous en retenons trois, qui nous ont aidé à cerner et construire notre définition de l'environnement musical.

Bregman [Bre94] tente de comprendre notre façon d'analyser une scène en n'utilisant que les sons et parle de scène sonore. Il s'interroge sur les processus mis en œuvre par l'être humain pour analyser, distinguer et interpréter les sons qui l'entoure, plus particulièrement dans le cas où plusieurs sons cohabitent. Il transpose au domaine audio, et en particulier à la musique, les modèles construits pour expliquer les processus d'analyse visuelle d'une scène. Ces processus sont basés sur des notions de proximité. Dans ce domaine, ce sont par exemple des couleurs similaires, des formes similaires, ou encore des objets proches spatialement. En musique, il distingue trois processus : le timbre, la proximité temporelle, et la proximité fréquentielle. Pour reconnaître une mélodie, les trois notions sont utilisées, pour un accord, il s'agit de la proximité fréquentielle.

Sundaram et Chang [SC00] définissent, quant à eux, la notion de scène sonore de la façon suivante : « une scène est une collection de sources sonores ; cette scène est dominée par quelques sources, qui sont supposées posséder des propriétés de stationnarité ». Ceci est reformulé synthétiquement par Cai *et al.* [CLC05] : « une scène sonore est un segment sonore consistant au niveau sémantique, qui est caractérisé par quelques sources dominantes ». De cette définition, nous déduisons celle de l'environnement musical.

Ainsi, l'environnement musical est la contribution de la musique à l'environnement sonore. Dans les documents audiovisuels, la musique intervient principalement dans trois cas : les extraits de musique seule (ex : clips, pauses musicales), la musique de fond (ex : titres du journal sur fond de musique, musique d'ambiance dans un film), et les ponctuations sonores (quelques secondes).

L'importance de la musique dans notre perception de la situation qui nous entoure se vérifie au travers de nombreux travaux en psychologie de la musique[6]. De nombreux journaux scientifiques [MP, PoM, MS, JNM, EMP] y sont consacrés.

La musique dans les documents audiovisuels joue un rôle, elle est le reflet d'une intention, elle crée une ambiance, un environnement musical, un temps de réflexion, une pause... De manière simplifiée, elle intervient principalement sous trois formes : les jingles, la musique de fond et les extraits musicaux.

[6]La psychologie de la musique cherche à « expliquer et comprendre le comportement musical et l'expérience musicale », ainsi que la définit l'article de la Wikipédia anglophone [Wik].

Au cours de ce chapitre, nous présentons quelques recherches menées pour caractériser l'environnement musical sous ses trois formes, les jingles, la musique de fond et les extraits musicaux. Tout d'abord, nous réunissons, dans la partie 2.2, une description des paramètres et outils de modélisation et classification couramment utilisés dans l'analyse de la musique – éléments que nous retrouvons dans les parties suivantes. La partie 2.3 est consacrée aux recherches menées sur les jingles. Puis, dans la partie 2.4, nous présentons les travaux réalisés pour la détection de la musique de fond. Enfin, dans la partie 2.5, nous décrivons les différentes approches utilisées pour caractériser les extraits musicaux. Nous nous attardons plus particulièrement sur ce type de musique, qui est au cœur de nos recherches. Plusieurs de ces approches sont plus aisées en contexte monophonique qu'en contexte polyphonique. De plus, la détection de chant peut être utile, voir même nécessaire pour certaines tâches.

2.2 Les outils du traitement automatique de la musique

Les paramètres communément utilisés pour la description de la musique sont soit empruntés à d'autres domaines, soit développés spécifiquement pour la musique. Pour chaque paramètre, nous décrivons rapidement son calcul, sa signification physique, et donnons, dans les cas où cela est possible, une traduction des termes anglophones. Le calcul est plus détaillé pour les paramètres les moins couramment utilisés.

2.2.1 Les paramètres « traditionnels »

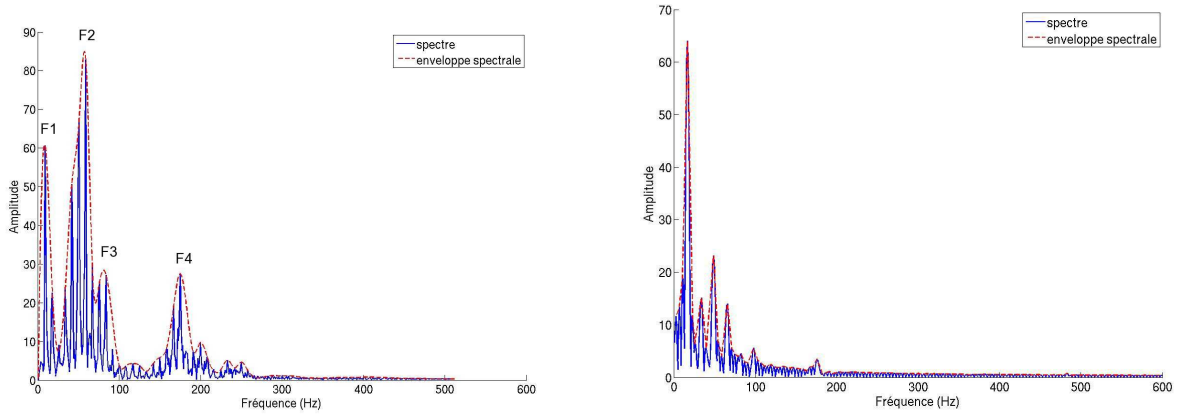
Le signal « musique » est un signal acoustique perçu par l'homme. Dans cette perception, une caractéristique forte est la hauteur du son. Cette perception a induit naturellement une utilisation de la transformée de Fourier et de ses dérivées.

L'introduction d'une modélisation du signal conduit à une deuxième série de paramètres : les MFCC et les paramètres de prédiction.

La Transformée de Fourier à Court Terme Il s'agit de la Transformée de Fourier réalisée sur une portion de signal (une « trame »), elle permet une analyse fréquentielle « instantanée » du signal. Une trame est typiquement d'une durée de 20 ms, pour assurer la quasi-stationnarité du signal sur la fenêtre d'analyse et une longueur perceptuelle significative.

L'enveloppe spectrale Il est intéressant de noter qu'il est difficile de trouver une définition pour ce terme si fréquemment utilisé en traitement du signal. Il s'avère en réalité que diverses définitions cohabitent, sans que la communauté scientifique puisse arriver à un consensus. Tout le problème réside dans la définition du terme « envelopper ».

Röebel *et al.* [RR05] proposent la définition suivante : « une fonction régulière qui passe par les sommets des partiels et suffisamment lisse pour ne pas modéliser les partiels ». L'enveloppe modélise donc la forme générale du spectre, sans tenir compte de la fréquence fondamentale. La figure 2.1 présente un exemple pour de la parole (a) et pour de la musique (b). Dans le cas de la parole, l'enveloppe spectrale rend compte des seuls formants.



(a) (b)

FIG. 2.1 – Enveloppe spectrale (a) d'une trame de parole (20 ms : $F1$, $F2$, $F3$ et $F4$ sont les formants, (b) d'une trame de musique (20 ms).

L'énergie par bande de fréquences Le spectre est divisé en bandes de fréquences, en nombre plus ou moins grand et réparties linéairement ou logarithmiquement. L'énergie est ensuite calculée dans chacune des bandes ainsi définies. Lorsque les bandes de fréquences sont réparties logarithmiquement, ce paramètre est parfois appelé **log frequency power coefficients**.

Plusieurs échelles logarithmiques existent, la plus utilisée dans le traitement automatique des sons est l'échelle perceptive **Mel**. Cette échelle, graduée en « Mels », est construite de telle façon qu'un écart constant en Mels correspondent à un écart constant en hauteur perçue [Cal89].

Les moments du spectre Seuls les quatre premiers moments du spectre sont utilisés : l'espérance (ou moyenne), la dispersion (ou variance), asymétrie (ou skewness) et le coefficient d'aplatissement (ou kurtosis).

Le centroïde spectral Le centre de gravité C du spectre est calculé par la formule suivante :

$$C = \frac{\sum_{n=0}^{N-1} S(\omega_n)\omega_n}{\sum_{n=0}^{N-1} S(\omega_n)} \tag{2.1}$$

où $S(\omega_n)$ est le module de la $n^{ième}$ composante fréquentielle ω_n, et N le nombre de classes fréquentielles.

Le Roll-Off C'est la fréquence en dessous de laquelle 80 % de l'énergie du spectre est contenue. Le centroïde spectral et le Roll-Off permettent de mesurer la répartition de l'énergie dans le spectre.

Le flux spectral Il s'agit de la distance euclidienne entre deux transformées de Fourier calculées sur des trames successives. Le flux spectral mesure les variations à court terme du spectre et sa dynamique.

Le constant-Q spectrum Cette transformée de type Fourier a été proposée par Brown [Bro91a]. Elle est calculée de la manière suivante :

$$X(k) = \frac{1}{N(k)} \sum_{n=0}^{N(k)-1} W(k,n)\ x(n) e^{-i2\pi Q \frac{n}{N(k)}} \tag{2.2}$$

où W est la fenêtre d'apodisation de longueur $N(k)$ et x le signal.

La taille $N(k)$ de la fenêtre d'analyse dépend de la fréquence analysée, ce qui aboutit à un pas d'échantillonnage fréquentiel non linéaire : le rapport Q entre deux indices fréquentiels successifs est maintenu constant. Cette représentation diffère des « log frequency power coefficients », en cela qu'elle donne un résultat intrinsèquement logarithmiquement espacé en fréquences. Le tableau 2.1 résume les différences entre le constant-Q spectrum et la transformée de Fourier.

Tab. 2.1 – Différences entre le Constant Q Spectrum (CQS) et la Transformée de Fourier Discrète (TFD) (avec F_e la fréquence d'échantillonnage).

	CQS	TFD
Fréquences	exponentielles : $(2^{1/24})^k f_{min}$	linéaires : $k\Delta f$
Fenêtre	variable : $N(k) = \frac{F_e Q}{f_k}$	fixe : N
Δf	variable : $f.k/Q$	fixe : F_e/N
$\frac{f}{\Delta f}$	fixe : Q	variable : k

Le ZCR Le Taux de Passage à Zéro (Zero-Crossing Rate ou ZCR en anglais) est le nombre de fois où le signal passe par la valeur zéro, sur une fenêtre de taille fixée (typiquement 10 ms). Il est directement lié à la fréquence du signal, si fréquence fondamentale il y a, mais il est sensible au bruit. Pour un signal bruité, ce paramètre sera élevé.

Les Mel Frequency Cepstal Coefficient - MFCC Paramètres « phare » du traitement de la parole, les MFCC, également appelés couramment coefficients cepstraux, sont probablement utilisés dans tous les domaines de l'analyse de l'audio. Les MFCC ont la propriété extrêmement intéressante de transformer un produit de convolution en une somme. Ainsi, si on considère que la voix est le produit de convolution entre une source harmonique (les cordes vocales) et un filtre (le conduit vocal), ils permettent de séparer la source du conduit. Il peut en être de même pour la musique. Ils sont répartis selon l'échelle Mel, pour refléter au mieux la perception humaine des sons.

Le schéma 2.2 résume le calcul des MFCC [Cal89].

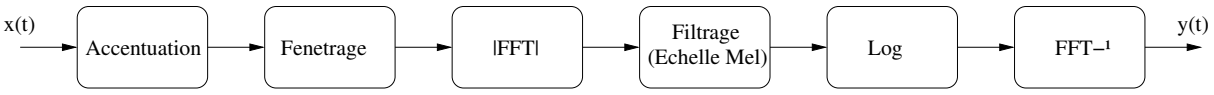


Fig. 2.2 – Calcul des MFCC.

Ces paramètres sont principalement utilisés en traitement de la parole, mais ils le sont également en musique.

Les paramètres de prédiction Dans cette catégorie, nous incluons deux types de paramètres : les Linear Predictive Coding (LPC) et les Wrapped Linear Prediction Coefficients (WLPC).

En codage par prédiction linéaire, le signal est modélisé par un modèle auto-régréssif gaussien. En parole, les pôles du modèle représentent les « formants », ou « fréquences de résonance » du conduit vocal. Les paramètres LPC sont les coefficients de ce filtre.

Il est classique, par transformée inverse du spectre, d'en déduire les LPCC : coefficients cepstraux issus d'une analyse par codage prédictif.

Les Wrapped Linear Prediction Coefficients sont une variante des coefficients LPC, dans laquelle est ajoutée un paramètre qui détermine la résolution fréquentielle du résultat.

Cette représentation du signal est très utilisée, pour l'analyse, le codage et la synthèse de la parole.

2.2.2 Les paramètres « musicaux »

Alors que les paramètres présentés ci-dessus sont couramment utilisés en analyse numérique de l'audio, qu'il s'agisse de parole ou de musique, les paramètres décrits dans la suite sont étroitement liés à la description d'un signal de musique « harmonique ».

Le coefficient harmonique Ce paramètre proposé par Cho *et al.* [CKK98] mesure le poids de la plus importante série dans une décomposition en somme de séries harmoniques. Son calcul se base sur l'autocorrélation dans le domaine temporel et dans le domaine fréquentiel :

– Calcul de l'autocorrélation temporelle R^T :

$$R^T(\tau) = \frac{\sum_{n=0}^{N-\tau-1} [\tilde{s}(n) \cdot \tilde{s}(n+\tau)]}{\sqrt{\sum_{n=0}^{N-\tau-1} \tilde{s}^2(n) \cdot \sum_{n=0}^{N-\tau-1} \tilde{s}^2(n+\tau)}} \tag{2.3}$$

avec s le signal à analyser et $\tilde{s}$ sa version centrée en zéro.

– Calcul de l'autocorrélation fréquentielle R^F :

$$R^F(\omega_\tau) = \frac{\sum_{\omega=0}^{N-\omega_\tau-1} \left[\tilde{S}(\omega) \cdot \tilde{S}(\omega+\omega_\tau)\right]}{\sqrt{\sum_{\omega=0}^{N-\omega_\tau-1} \tilde{S}^2(\omega) \cdot \sum_{\omega=0}^{N-\omega_\tau-1} \tilde{S}^2(\omega+\omega_\tau)}} \tag{2.4}$$

avec S le module de la transformée de Fourier de s, $\tilde{S}$ sa version centrée en zéro et $\omega_\tau = 2\pi N/\tau$.

– Combinaison des deux autocorrélations :

$$R(\tau) = \beta \cdot R^T(\tau) + (1-\beta) R^F(\tau) \tag{2.5}$$

– Le coefficient harmonique H_a est alors défini de la manière suivante :

$$H_a = \max_\tau R(\tau) \tag{2.6}$$

Si le coefficient harmonique est faible, cela signifie qu'aucune série harmonique n'est importante, il s'agit de bruit. Une valeur élevée du coefficient harmonique s'interprète par l'existence d'une ou plusieurs séries harmoniques de forte importance dans le signal.

Une représentation chromatique du signal : le chroma vector Ce paramètre a pour but de représenter l'importance de chacune des 12 notes de la gamme dans un accord donné. Il s'agit d'une représentation spectrale particulière, en dimension 12, issue de la proposition de Shepard [She64] de représenter les fréquences sous une forme circulaire.

Le spectre est divisé en 12 classes, correspondant aux 12 demi-tons de la gamme. Le $i^{ième}$ coefficient du chroma vector correspond à l'énergie dans les bandes de fréquences correspondant à la $i^{ième}$ note de la gamme, à toutes les octaves possibles. Le vecteur résultant est ensuite parfois normalisé.

Dans la suite, nous noterons $C(t) = [c_{Do}(t), c_{Do\sharp}(t)...c_{Si}(t)]$ le chroma vecteur associé à la trame t.

Les « Key profile » Les « profils de clé » modélisent l'importance de chaque demi-ton dans une gamme. En se basant sur la définition de la gamme (le profil « diatonique »), à une note de la gamme est attribuée la valeur 1, 0 aux autres. Cette première solution, un peu brute, interdit toute note accidentellement altérée. Deux autres possibilités sont d'utiliser les statistiques proposées par Krumhansl [Kru90] ou par Temperley [Tem01] ; ils décrivent les fréquences d'apparition des différents demi-tons dans une gamme, d'un point de vue expérimental. Les trois profils (diatonique, de Krumhansl et de Temperley), sont résumés, pour les gammes Majeure et mineure, dans le tableau 2.2, le demi-ton n° 0 correspond à la tonique (la première note de la gamme).

TAB. 2.2 – Comparaisons des Key profile.

n° du demi-ton	K M	K m	T M	T m	Dia M	Dia m
0	6.35	6.33	5.0	5.0	1	1
1	2.23	2.68	2.0	2.0	0	0
2	3.48	3.52	3.5	3.5	1	1
3	2.33	5.38	2.0	4.5	0	1
4	4.38	2.6	4.5	2.0	1	0
5	4.09	3.53	4.0	4.0	1	1
6	2.52	2.54	2.0	2.0	0	0
7	5.19	4.75	4.5	4.5	1	1
8	2.39	3.98	2.0	3.5	0	0
9	3.66	2.69	3.5	2.0	1	1
10	2.29	3.34	1.5	1.5	0	1
11	2.88	3.17	4.0	4.0	1	0

K : Profil de Krumhansl
T : Profil de Temperley
Dia : Profil diatonique
M : Majeur
m : mineur harmonique

Une représentation perceptuelle : les cercles des quintes, des tierces Majeures et des tierces mineures Ces outils permettent de représenter les proximités perceptuelles des différentes tonalités : deux tonalités majeures sont perceptuellement proches si elles sont distantes d'une quinte (ce sont les tons « voisins »). Le cercle des quintes a initialement été décrit par Krumhansl [Kru90] (figure 2.3). On peut alors projeter un vecteur (chroma vecteur par exemple) $\overrightarrow{v} = [v_{DO}, v_{DO\sharp}...v_{SI}]$ dans le cercle des quintes de la manière suivante : $COF(\overrightarrow{v}) = \sum_{i \in Do, Do\sharp..Si} v_i.\overrightarrow{u_i}$ avec $\{\overrightarrow{u_i}\}_i$ les vecteurs unité de chacune des clés.

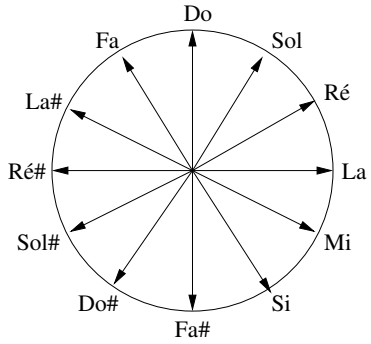


FIG. 2.3 – Le cercle des quintes, avec les vecteurs unité de chacune des clés.

Un exemple de projection d'un chroma vecteur est donné sur la figure 2.4 : la contribution de chacune des notes est indiquée par un vecteur plein noir, le vecteur $\overrightarrow{COF}$ résultant est le vecteur pointillé en gros trait. Ici, l'accord joué est un accord parfait de Do Majeur. Les notes qui contribuent le plus sont, dans l'ordre, le Do, le Sol et le Mi. Les autres notes sont anecdotiques.

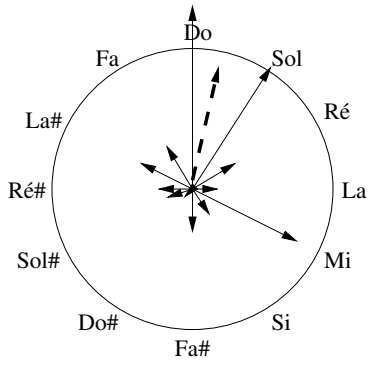


FIG. 2.4 – Un exemple de projection d'un accord (Do M) dans le cercle des quintes.

De la même façon, on peut construire le cercle des tierces majeures et le cercle des tierces mineures (figure 2.5).

En ajoutant des vecteurs unité comme dans le cercle des quintes, il est également possible de projeter un vecteur dans chacun de ces deux cercles.

Le centroïde tonal En projetant un chroma vector dans chacun de ces trois cercles (quintes, tierces Majeures et tierces mineures), Harte *et al.* [HSG06] construisent un vecteur à 6 dimensions (deux coordonnées dans chaque cercle), qu'ils appellent le centroïde tonal.

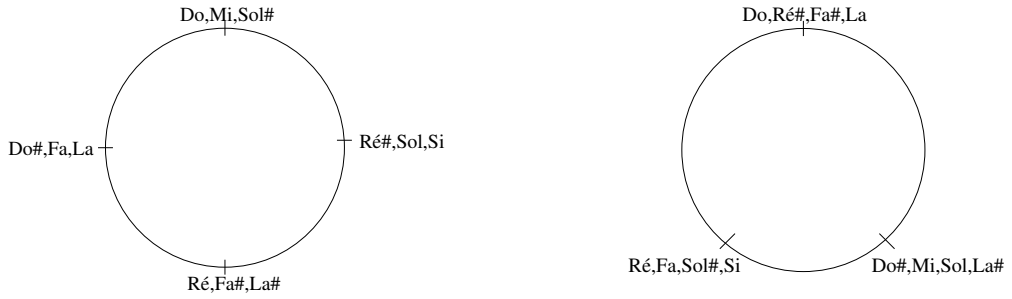


FIG. 2.5 – Cercles des tierces (a) mineures et (b) Majeures.

2.2.3 Les méthodes de classification et de modélisation

Les études rapportées dans ce document relèvent principalement du domaine de la reconnaissance des formes. Les méthodes de classification utilisées sont des outils classiques dérivant des approches génératives (modélisation probabiliste essentiellement), et des approches discriminantes. Nous rappelons brièvement leurs principes.

Les Modèles de Markov Cachés Les Modèles de Markov Cachés (Hidden Markov Models ou HMM en anglais) permettent de modéliser des enchaînements temporels au travers d'états, en tenant compte de leur durée. Cette méthode, développée dans les années 1965-1970 par Leonard E. Baum [Bau72] a été utilisée pour la reconnaissance de la parole ; elle est idéale pour modéliser chaque mot comme un enchaînement de sons de longueur variable.

De la même façon, ils peuvent être utilisés pour modéliser n'importe quelle séquence temporelle qui respecte une certaine grammaire. En musique, ils sont utilisés par exemple pour modéliser les enchaînements d'accords [MDH+07, LS08] ou de tonalités [IK09]. Les tonalités, tout comme les accords au sein d'une tonalité, s'enchaînent en respectant certaines règles. Pour construire les HMM, on peut ainsi, soit utiliser les nombreuses théories musicales, soit utiliser des corpora annotés manuellement. Dans ce cas, il faut estimer les probabilités de transitions entre états à l'aide des algorithmes classiques d'apprentissage des HMM, pour rendre compte de ces règles.

À titre d'exemple, citons la reconnaissance d'accords : à partir d'une suite d'observation de type « chroma vector », l'étape de classification/reconnaissance se fait en cherchant, dans le HMM représentant le modèle, le chemin qui maximise la probabilité d'observation de cette suite.

Les Modèles de Mélanges de Gaussiennes Les Modèles de Mélanges de Gaussiennes (Gaussian Mixture Models ou GMM en anglais) sont des lois de probabilité uni- ou multi-dimensionnelles, très utilisées pour modéliser des répartitions inconnues ou susceptibles de présenter plusieurs modes dont le nombre est souvent inconnu. Ces mélanges sont

couramment utilisés pour modéliser la voix d'un locuteur, ou la variabilité d'un son en parole. De fait, elles présentent l'avantage d'avoir été très étudiées ; leurs propriétés sont connues, les algorithmes d'estimation des paramètres sont éprouvés...

Un modèle de mélange de Gaussiennes est la somme pondérée de N Gaussiennes uni- ou multi-dimensionnelles. Pour un nombre de Gaussiennes N fixé, la loi de probabilité de la distribution est définie de la manière suivante :

$$g(x, \mu_1, \Sigma_1, \mu_2, \Sigma_2, ..., \mu_N, \Sigma_N) = \sum_{k=1}^{N} \pi_k f(x, \mu_k, \Sigma_k) \tag{2.7}$$

avec $f(x, \mu_k, \Sigma_k)$ la loi normale uni- (resp. multi-) dimensionnelle de moyenne (resp. vecteur de moyennes) μ_k et de variance (resp. matrice de covariance) Σ_k et π_k le poids de la $k^{ième}$ composante. Théoriquement, ces mélanges permettent d'approcher nombre de distributions probabilistes, pourvu que le nombre de composantes soit suffisant.

Dans un problème à M classes, la distribution des paramètres pour chaque classe est modélisée par un GMM, le processus de décision se fait ensuite par la méthode du maximum de vraisemblance [DHS01].

Les Machines à Vecteur de Support Les Machines à Vecteur de Support [BGV92] (Support Vector Machine, ou SVM en anglais) sont des outils de classification discriminants développés pour les problèmes à deux classes. Dans le cas de données séparables linéairement dans un espace à N dimensions, les deux classes sont séparables par un hyperplan. Ceci consiste à rechercher le meilleur hyperplan H, c'est-à-dire celui qui maximise sa distance d aux frontières de chaque classe. Tout l'intérêt des SVM est que cet hyperplan optimal peut être caractérisé par les points de chaque classe qui en sont les plus proches : les « vecteurs de support » (figure 2.6).

Dans le cas où les données ne sont pas séparables dans l'espace de représentation, elles sont projetées dans un espace de dimension supérieure où elles sont séparables. Le problème est évidemment de trouver le bon espace image, au travers d'une fonction noyau adéquate qui correspond au produit scalaire dans ce nouvel espace.

Des méthodes ont été développées pour étendre cet outil aux problèmes à plusieurs classes :

- La méthode « un contre un » : $N(N-1)$ SVM bi-classe sont créés, pour apprendre toutes les séparations existant entre chaque couple de classes.
- La méthode « un contre tous » : N SVM bi-classe sont créés, pour apprendre les frontières de chacune des classes.

Les k-Plus Proches Voisins Les k-Plus Proches Voisins (k-PPV, k-Nearest Neighbor ou k-NN en anglais) sont un algorithme de classification qui permet d'ignorer la distribution probabiliste des données. Il est basé sur une estimation locale de la densité de

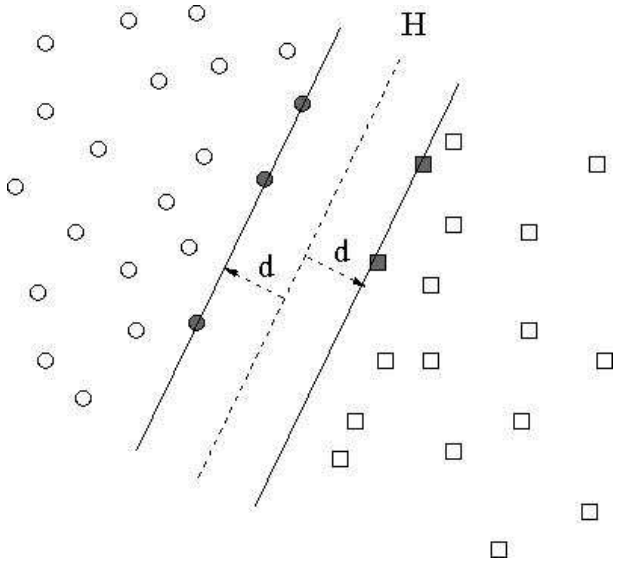


FIG. 2.6 – Schéma explicatif de la méthode SVM. H est l'hyperplan séparateur. Les vecteurs support sont grisés.

probabilité pour chaque classe. À partir d'un ensemble d'apprentissage composé d'observations vectorielles, chacune étant associée à une classe, la décision se prend, pour chaque nouvelle observation, par vote majoritaire sur les classes des k observations de l'ensemble d'apprentissage les plus proches. La mise en œuvre de cet algorithme nécessite de régler la variable k.

Il est certain que l'avantage de cet algorithme est de prendre en compte un nombre quelconque de classes (contrairement à la technique SVM), et de n'avoir aucun *a priori* sur les lois probabilistes. En revanche, le coût calculatoire est relativement élevé, même si des algorithmes ont été développés pour ne pas avoir à calculer la distance à tous les vecteurs de la base d'apprentissage, mais seulement à une partie d'entre eux, ce qui réduit considérablement le coût de calcul [IM98].

Les Réseaux de Neurones Les Réseaux de Neurones [Koh84] (Neural Networks en anglais) les plus employés en classification automatique sont les Perceptrons Multicouches (Multilayer Perceptron).

Le neurone formel (figure 2.7) est connu pour résoudre un problème de classification à deux classes. Le perceptron multicouches (figure 2.8) permet de réaliser une classification en k classes de données non séparables linéairement. Les sorties de chaque neurone de sortie peuvent être interprétées comme des valeurs prédictives, ou assimilées à des scores probabilistes. De ce fait, le perceptron est mis en œuvre pour fusionner des décisions issues de plusieurs classifieurs primaires.

Un autre intérêt du perceptron est de disposer d'algorithmes d'apprentissage efficaces (par exemple l'algorithme de rétropropagation du gradient [KS96]), mais un inconvénient

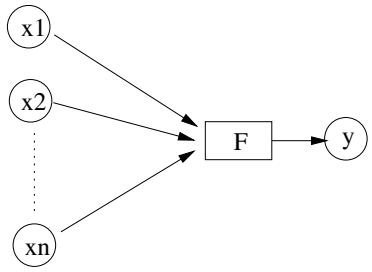


FIG. 2.7 – Neurone formel.

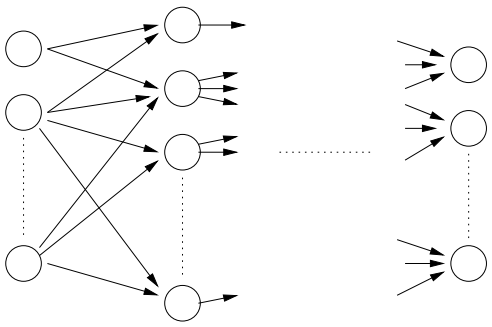


FIG. 2.8 – Structure d'un réseau de neurones. F est une fonction non linéaire : une fonction « seuil » ou de type tangente hyperbolique.

réside dans la difficulté de définir sa topologie (nombre de couches, nombre de neurones par couche).

2.2.4 Les librairies de calcul

De nombreuses librairies de calcul sont disponibles en ligne pour les diverses méthodes de classification que nous avons présentées ci-dessus.

Elles sont toutes disponibles dans le langage de programmation Matlab[7], chacune étant proposée avec des options permettant d'utiliser tout l'éventail des possibilités de la méthode.

Des alternatives libres existent, codées en C pour la plupart. La librairie TORCH[8] propose des programmes pour toutes ces méthodes (GMM, HMM, SVM, k-PPV, réseaux de neurones), et même d'autres.

Pour les SVM, on pourra également utiliser *libsvm*, dont les programmes sont disponibles en ligne[9]. Cette librairie permet, outre l'apprentissage et la classification, d'utiliser des fonctions noyaux, et la recherche des paramètres optimaux.

[7] www.mathworks.fr/products/matlab/
[8] http://www.torch.ch/
[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Pour les réseaux de neurones, le logiciel SNNS[10] (Stuttgard Neural Network Simulator), proposé par l'université de Stuttgard, est très complet.

Enfin, pour les modèles de Markov cachés, la librairie HTK[11] est peut-être la plus utilisée.

2.3 Les jingles

2.3.1 Définition

Selon l'Office Québécois de la langue française[12], les jingles sont, au sens propre, les « ritournelles publicitaires ». Dans notre travail, nous utilisons la définition qu'utilise Anne-Marie Gustave dans son article en ligne « La loi des jingles » [Gus08] : « [...] virgules, tourne pages, indicatifs, fins de titre, tapis déroulants ou auto-promos, [...] autant d'éléments sonores qui constituent l'habillage d'une radio. Son identité. Sa ligne esthétique ». Dans cette définition, il est clair d'une part qu'il y a une forte ressemblance entre les jingles d'une même radio (l'identité de la radio), et d'autre part qu'il y a une forte dissemblance entre les jingles de radios différentes (il ne faudrait pas que l'auditeur confonde deux radios !).

Les jingles sont donc un moyen efficace d'identifier une radio, mais ils permettent également, au sein d'une radio, d'identifier les différentes émissions.

2.3.2 Les caractéristiques des jingles

Les caractéristiques des jingles (durée, place dans le flux audio, présence ou non de parole superposée à la musique) sont différentes selon le rôle qu'ils jouent.

Tout d'abord, ils proposent des pauses, à la manière de la virgule à l'écrit, ils sont alors appelés « virgules », nous les appelons « **jingles courts** ». Les jingles courts sont *d'une durée inférieure à 5 secondes*, et *ne contiennent habituellement pas de parole*. Ils sont présents *au milieu* d'une émission.

Les jingles peuvent également annoncer le début ou la fin d'une séquence ou d'une émission, ce sont les **génériques ou indicatifs**. C'est par exemple le jingle d'annonce des informations ou de la météo. À l'inverse des précédents, ils sont placés *en début et/ou fin d'émission*. Dans ce cas, il y a *éventuellement de la parole superposée à la musique*, annonçant par exemple le titre de l'émission, ou le présentateur. Dans le cas d'un générique d'émission, la musique donne le ton de ce qui va suivre : les émissions de reportage ont plutôt en générique des musiques « haletantes », les émissions de divertissement des musiques « joyeuses »...

[10] http://www.ra.cs.uni-tuebingen.de/SNNS/
[11] http://htk.eng.cam.ac.uk/
[12] www.granddictionnaire.com

Enfin, chaque radio doit rappeler son nom régulièrement, ceci est généralement réalisé par un indicatif, qui sera appelé « **jingle** ». Il est intéressant de noter que ce dernier type indicatif *contiendra toujours de la parole* (le nom de la radio, éventuellement accompagné d'une autre information, l'heure par exemple : « France Inter, il est 8 heures »).

2.3.3 Quelques travaux réalisés sur le sujet

En résultat annexe de leurs travaux [BBP04], Brück *et al.* ont réalisé un détecteur de génériques. L'objectif était de définir une « signature audio », capable de caractériser entièrement et de façon unique un document audio, tout en étant relativement compacte. Cette signature est composée des huit énergies dans 8 bandes de fréquences (logarithmiquement réparties entre 300 et 2000 Hz), calculées toutes les 40 ms. Au final, la signature est effectivement compacte, puisqu'il y a un rapport 30 entre la durée d'un document et le temps de calcul de sa signature.

Pour retrouver un document, ou un extrait, dans une collection, il suffit alors de retrouver sa signature. Ceci est réalisé en deux étapes : tout d'abord, la dissimilarité entre sa signature et celles présentes dans la base de données est calculée. Cette dissimilarité est la distance euclidienne, filtre à filtre, entre les signatures. Une étape de lissage est ensuite réalisée pour nettoyer la courbe. Une application proposée pour tester cet algorithme a été la recherche de génériques d'émission.

Dans leurs travaux sur la structuration, Carrive *et al.* [CPR00] insistent sur le rôle structurant des jingles pour l'indexation, sans indiquer leur méthode de détection automatique.

Des travaux ont également été réalisés sur cette tâche au sein de notre équipe [PAO04], travaux que nous présentons dans la partie 5.1.3.

2.4 La musique de fond

La musique de fond est sûrement la plus difficile à détecter, et plus encore à caractériser. Elle est, par définition, recouverte par d'autres sons : de la parole (journaux d'information, films, vidéos personnelles), des bruits divers (films, vidéos personnelles), voire même les deux. Nous nous intéressons ici uniquement à la musique dans les films et vidéos personnelles.

Dans les documents audiovisuels, de nombreuses informations peuvent être extraites en ne se basant que sur l'image. Effectivement, la plupart des concepts actuellement extraits automatiquement le sont en utilisant l'image uniquement (voir par exemple les systèmes proposés dans les dernières campagnes d'évaluation TRECVideo [SOK06]). L'analyse de la bande audio a cependant un énorme avantage sur l'image : la quantité de données à analyser est nettement moins grande, le temps de traitement est donc *a priori* également moindre.

Jusqu'à aujourd'hui, très peu de travaux se sont encore intéressés à caractériser cette musique de fond. Nous les évoquons dans la partie 5.

Le premier problème réside dans sa détection. En effet, la musique de fond étant par nature recouverte d'un autre son, *a priori* plus fort, et les travaux étant réalisés sur des bandes son monocanal, une localisation temporelle précise des extraits de musique est difficile.

Historiquement, les premiers outils ont été développés pour rechercher les zones de parole (et non les zones de musique). En effet, cette détection est nécessaire comme pré-traitement pour la transcription de la parole. Cependant, comme il s'agissait de transcriptions de radio, voire même plus précisément de journaux radiophoniques (ESTER [GGM+05]), le problème s'est souvent ramené à distinguer la parole de la musique [SS97]

Récemment (depuis deux ou trois ans), des recherches se focalisent sur la détection de la musique, quels que soient les autres sons présents (parole, bruit, les deux ou aucun). Cette évolution des recherches est certainement due à la volonté d'analyser des films, dans lesquels la musique est à la fois omniprésente et très significative, et à l'explosion des vidéos amateurs, que ce soit sur internet ou pour des collections personnelles.

2.4.1 Les paramètres et les modélisations

Des systèmes développés pour des tâches de traitement de la parole (ou de traitement du son en général), nous retiendrons comme paramètres les MFCC, la modulation de l'énergie à 4 Hz, le centroïde spectral, le zero-crossing rate, le Roll-Off point, le flux spectral, le pourcentage de trames de basse énergie. Ces paramètres, proposés par Scheirer [SS97] et Tzanetakis [Tza04] ont été testées par Izumitani *et al.* [IMK08] spécifiquement pour la détection de musique de fond. Seyerlehner *et al.* [SPS07] ont également testé les LPC (Linear Predictive Coefficients), Giannakopoulos *et al.* [GPT08] utilisent enfin l'entropie de l'énergie et le taux de voisement.

Les autres paramètres utilisés pour cette tâche sont issus du traitement de la musique. Lee et Ellis [LE08] utilisent la fréquence fondamentale et le rythme. Giannakopoulos *et al.* [GPT08] basent leurs paramètres sur le chroma vector (voir partie 2.2).

La prise de décision est faite par des classifieurs classiques : des GMM, des SVM, des réseaux de neurones, des k-plus proches voisins... Une étape de lissage est parfois ensuite ajoutée, pour passer de l'échelle d'analyse (centiseconde) à une échelle de décision plus réaliste (une seconde). Ceci permet de tenir compte de la continuité temporelle de la musique.

2.4.2 Les performances

Les tests sont réalisés sur différents types de corpora : des extraits télévisés pris au hasard, en gardant malgré tout une grande diversité d'émissions (films, débats, documen-

taires, informations, talk-show...) [SPS07], des films [GPT08], des vidéos amateurs prises au hasard sur le web sur des sites de partage de vidéos (youtube[13], dailymotion[14], etc.), ou encore des données artificielles [IMK08]. Les données professionnelles (télévision ou films), sont acoustiquement plus propres que les données amateures. Les données artificielles permettent de mesurer assez précisément le Rapport Signal à Bruit (Signal to Noise Ratio ou SNR en anglais), et d'avoir donc une idée de l'influence de celui-ci sur les résultats.

Les résultats sur les données télévisuelles sont de l'ordre de 90 % d'accuracy[15] [SPS07]. Sur des données artificielles, Izumitani *et al.* obtiennent un taux d'erreur à l'échelle de la trame de 8 % [IMK08] (une trame d'analyse audio est typiquement de 20 ms). Sur les films, la F-mesure est de 90.5 % [GPT08].

Notons que, cette tâche étant très récente, ni les corpora, ni les mesures ne sont unifiés, rendant difficile une comparaison des performances des différentes méthodes.

2.5 Les extraits musicaux

Dans les flux audiovisuels, les extraits musicaux sont souvent caractéristiques de la nature du document, ou à tout le moins choisis selon un critère non aléatoire. Pour un film par exemple, le choix des musiques est déterminé par les situations. Leur caractérisation peut se faire à différents niveaux, et selon différents points de vue. Les données importantes sont naturellement le style, mais aussi des paramètres plus « musicologiques » :

- des paramètres donnés à l'échelle de l'extrait musical : l'effectif, la tonalité, la pulsation...,
- des paramètres donnés à l'échelle de la trame ou de la note : la transcription de la mélodie principale, de la partition, des accords, ou encore des paroles en présence de chant.

Nous présentons un bref état de l'art des différents travaux menés sur chacune de ces thématiques. Notons que la plupart des recherches sont faites sur de la musique occidentale, dans laquelle les notions de tonalité et de rythme sont bien définies. Dans chaque thématique, nous nous attachons plus précisément à définir la question étudiée, les termes techniques associés au sujet et les éventuels problèmes de fond rencontrés. Ensuite, nous présentons, si possible, un bref historique des travaux réalisés sur la question, et les connaissances et performances actuelles.

[13] http://www.youtube.com/

[14] http://www.dailymotion.com

[15] Nous utiliserons tout au long de cet ouvrage le terme anglais, couramment employé par l'ensemble de la communauté

2.5.1 L'effectif, le timbre

Les questions Quels sont les instruments présents ? Y a-t-il des chanteurs ? Si oui, quel est leur registre (soprano, alto, ténor, basse) ? Les deux premières questions sont généralement abordées séparément. Dans cette partie, nous ne traitons que de l'identification des instruments ; l'état de l'art pour la détection du chant est présenté dans la partie 4.2.

Terminologie L'identification des instruments est également appelée l'identification du timbre, le timbre est en effet caractéristique d'un instrument. L'ANSI [Ins60] (American National Standards Institute) définit le timbre comme « [la] caractéristique sonore, qui permet à un auditeur de juger que deux sons présentés de la même façon, de même hauteur, durée et intensité sont différents ». L'ANSI précise que « [le timbre] dépend principalement du spectre, mais aussi de la forme d'onde, de la pression acoustique, de la disposition des fréquences à l'intérieur du spectre, et des caractéristiques temporelles du stimulus ». Il apparaît donc que le timbre, s'il est évident à notre oreille, est une notion encore mal comprise au niveau physique. Pour plus de détails sur le timbre, ses définitions et caractéristiques, on pourra se reporter par exemple à la thèse de Marozeau [Mar04c].

Historique Le problème de l'identification des instruments (et indirectement de la reconnaissance du timbre) a d'abord été étudié pour des instruments solo, jouant une note [FAP99, Bro99, EK56, KC05, LR06], puis sur des instruments solo jouant une phrase musicale [LR04]. Les instruments sont évidemment plus faciles à reconnaître lorsqu'ils jouent en solo que dans une polyphonie. Pour un instrument donné, une grande partie de l'information sur son identité se trouve dans l'attaque des notes, ce qui explique que le problème ait au départ été traité en considérant des notes isolées avant de considérer des phrases musicales.

En contexte monophonique, les premiers paramètres utilisés ont été la pente de l'attaque, le degré de synchronisme des harmoniques, les MFCC [Bro99, EK56], l'énergie dans les hautes fréquences [FAP99], le Constant-Q spectrum [Bro99], la modulation de l'énergie, le centroïde spectral, diverses mesures statistiques calculées sur le spectre [EK56], et le vibrato[16][KC05].

Actuellement, tout en continuant les recherches sur les sujets précédents, de nouveaux travaux, que nous présentons ci-dessous, se penchent sur l'identification d'instruments en contexte polyphonique [VR05, ERB05, ER06, MBTL07, KGK+07].

Méthodes actuelles

Les paramètres extraits du signal restent les paramètres classiques utilisés en traitement

[16] Le vibrato est un paramètre défini à partir de la fréquence fondamentale. Pour une description complète de ce paramètre, on se reportera à la partie 4.3.1

automatique de l'audio : paramètres temporels, MFCC, LPC et leurs dérivées, et paramètres fréquentiels (un accent particulier est mis sur ces derniers, puisque le timbre est sensé être de nature spectrale). Quelques paramètres sont spécifiques à cette tâche, comme l'énergie par octave, un cas particulier de l'énergie par bande de fréquences [ER06]. Une étude très complète des performances de plus d'une centaine de paramètres a été menée par Every [Eve08]. Au cours de plusieurs expériences, il mesure l'importance de chaque paramètre pour la reconnaissance des instruments. Il en ressort que la fréquence fondamentale est, justement, un paramètre fondamental, puisqu'elle est toujours le paramètre le plus discriminant. Viennent ensuite les moments spectraux et harmoniques. Enfin, Livshin et Rodet [LR06] ont mesuré l'information contenue dans la partie non harmonique, et ont conclu que l'information est principalement contenue dans la partie harmonique, sauf pour quelques instruments tels que la clarinette, la flûte et la trompette, pour lesquels on entend le souffle ou des « clics ».

La classification est réalisée à l'aide d'outils classiques : GMM, SVM,... Une attention particulière est généralement prêtée à l'apprentissage des modèles. En effet, en contexte polyphonique, il est difficile d'obtenir des exemples pour toutes les combinaisons possibles d'instruments. Deux cas de figure sont envisageables :

- *Le nombre d'instruments à reconnaître est relativement restreint* (jazz, rock par exemple), et modéliser chacune des combinaisons est possible [ERB05, MBTL07]. Les modèles sont appris soit à partir d'extraits solo mixés artificiellement entre eux, soit à partir d'extraits polyphoniques. Dans ce cas, pour le test, l'exemple est comparé à chacun des modèles pour en identifier l'effectif.
- *Le nombre d'instruments à reconnaître est très important* et il est alors difficilement envisageable de créer un modèle pour chacune des combinaisons. Il peut alors être judicieux de n'avoir qu'un modèle par instrument, et de ne pas avoir les modèles des combinaisons, pour une question de temps de calcul. Dans ce cas, la méthode est la suivante [MBTL07] : une première étape de séparation de sources permet d'obtenir les instruments séparés, qu'il ne reste ensuite plus qu'à classer en utilisant les méthodes éprouvées pour les instruments solo.

Notons qu'une information *a priori* est parfois ajoutée sur la composition du signal. Essid *et al.* [ERB05, ER06] connaissent le style et peuvent en déduire les diverses instrumentations possibles. Vincent et Rodet [VR05] utilisent une information sur le nombre d'instruments présents.

Performances Pour la reconnaissance d'instruments isolés, les résultats sont de l'ordre de 93 % d'accuracy pour 19 instruments à reconnaître [KC05], ou encore 97 % pour la reconnaissance des grandes classes instrumentales : cordes pincées, cordes frappées, cordes frottées, bois, hanches simples, hanches doubles, cuivres, idiophones[17].

[17] Un idiophone est un instrument dont le matériau lui-même produit le son lors d'un impact, par exemple le xylophone

La reconnaissance d'instruments en contexte polyphonique est évidemment une tâche beaucoup plus difficile. Essid *et al.* [ERB05, ER06] atteignent une accuracy de 53 % en s'intéressant à un style particulier : le jazz. Martins *et al.* [MBTL07] obtiennent une précision de 64 % et un rappel de 56 % sur des extraits de deux à quatre notes mixées artificiellement, pour six instruments possibles. Cependant, les auteurs notent que les performances chutent drastiquement avec l'augmentation du nombre d'instruments. Ce fait est confirmé par Kitahara *et al.* [KGK+07], pour qui le taux de reconnaissance, de 53,4 % pour des duos, passe à 46,5 % pour des quartets, pour cinq instruments possibles.

Il convient de noter que certains instruments posent problème, notamment les percussions, à cause du caractère non harmonique du son qu'ils produisent.

2.5.2 La tonalité

La question Quelle est la tonalité de l'extrait musical, parmi les 24 tonalités possibles ? Avec cette formulation de la question, on s'intéresse évidemment à des styles musicaux qui suivent cette classification : il s'agit de la musique occidentale et tonale. Les 24 tonalités possibles correspondent aux 12 demi-tons de la gamme, avec les variantes Majeure et mineure pour chacun. On est dans un problème de classification fermée.

Historique Krumhansl [Kru90] a été l'un des premiers à s'intéresser à l'identification de la tonalité. Dans ces travaux, il analyse des partitions, et se base sur les « Key profile » qu'il a définis (voir dans la partie 2.2.2) pour déterminer la tonalité d'un extrait. Notons que dans les premières recherches, seuls des extraits musicaux qui ne changeaient pas de tonalité ont été considérés. Les recherches actuelles envisagent les modulations : la tonalité peut changer au cours du morceau.

Méthodes actuelles
La plupart des méthodes utilisent les « chroma vectors » comme paramétrisation. Elles cherchent ensuite à savoir si les notes présentes correspondent à une tonalité particulière. Pour cela, la méthode la plus simple est d'apprendre des modèles pour la répartition des chroma vectors pour chacune des tonalités. On pourra par exemple utiliser les « Key profile ». Pour estimer la tonalité d'un extrait donné, Gómez compare la moyenne des chroma vectors [Gom06] à chacun des modèles. İzmirli [Izm05] compare lui chaque chroma vector à chacun des modèles, et décide que la tonalité générale de l'extrait est celle qui a le plus grand score en moyenne.

Inoshita et Katto, [IK09] introduisent le vecteur de tonalité K (« tonality vector »), de dimension 12, qu'ils projettent dans le cercle des quintes. Cette projection mesure l'adéquation entre les notes présentes et une tonalité donnée. Ce paramètre est construit à partir d'un chroma vector $C(t) = [c_{Do}(t), c_{Do\sharp}(t)...c_{Si}(t)]$. Chaque composante du vecteur de tonalité correspond à une gamme, et est la somme des contributions de chacune des

notes de la gamme considérée. Ainsi, la composante de $K(t)$, correspondant à la tonalité Do M (ou La m), est la somme des composantes de $C(t)$ correspondant aux notes Do, Ré, Mi, Fa, Sol, La et Si :

$$K(t) = [k_{Do}(t), k_{Do\sharp}(t)...k_{Si}(t)] \tag{2.8}$$

où

$$k_{P_n}(t) = \sum_{i=0}^{11} w_i f(c_{P_{i+n(mod\ 12)}}(t)),\ \text{avec } P_0 = Do, ..., P_{11} = Si \tag{2.9}$$

avec w_i un poids associé à chaque note, poids choisi en se basant par exemple sur un des « Key profile », et f une fonction de lissage de la puissance des notes.

Un HMM permet de suivre la tonalité, et d'en repérer les changements. On obtient ainsi un nuage de points qui doit normalement être dans le quadrant du cercle correspondant à la tonalité du morceau. Dans le cas d'une modulation, plusieurs nuages de points sont obtenus. Il est à noter que cette méthode, si elle est relativement efficace avec 70 % de bonnes réponses, ne permet pas de distinguer les gammes relatives majeure et mineure, qui utilisent les mêmes notes.

Cette proposition est semblable au « keyogramme » que proposent Hart *et al.* [HFC07], où est calculée, pour chaque instant et pour chaque gamme possible, la contribution des notes de la gamme. Cependant, cette méthode ne permet pas non plus de distinguer les gammes relatives.

Une autre possibilité enfin est de se baser sur l'analyse des accords, ou mieux encore de la suite d'accords. Les enchaînements d'accords (les cadences notamment), sont des caractéristiques très fortes de la tonalité. Comme le disent Mauch *et al.* [MDH+07], « [La tonalité] est implicitement codée dans les séquences d'accords suffisamment longues ». Il est même très souvent possible de déterminer la tonalité d'un morceau en ne se basant que sur la lecture du (des) dernier(s) accord(s) d'une partition. Ainsi, Noland et Sandler [NS06] modélisent les changements possibles de tonalité par un Modèle de Markov Caché. Les probabilités de transitions entre états sont les probabilités de changer de tonalité. Les observations du modèle sont soit des transitions entre accords, soit des paires d'accords (par exemple, une transition Sol M - Do M correspond très probablement à une tonalité de Do M). Les différentes probabilités de transitions entre accords, pour chacune des clés, ont été apprises sur des séquences d'accords manuellement annotées. Cette méthode a l'avantage de distinguer les gammes majeures de leurs relatives mineures.

Performances Les expériences menées par chaque équipe l'ont été sur des corpora différents, incluant soit uniquement de la musique classique [Pee06, Izm05, Pau04], soit de la musique classique et de la musique pop [IK09], soit uniquement de la musique pop [NS06], soit encore de la musique traditionnelle [HFC07]. Ainsi, les performances sont difficilement comparables. Actuellement, les meilleurs performances sont de l'ordre de 85

à 90 % de bonne identification de la tonalité. Pauws [Pau04] remarque qu'en considérant comme équivalentes les tonalités relatives (Do M et La m) et les tons voisins (la quinte : Sol M et Mi m et la quarte : Fa M et Ré m), ses résultats s'améliorent de près de 20 %, passant de 75 % à 94 % de bonne identification. Les paramétrisations du signal semblent pertinentes d'un point de vue musicologique.

2.5.3 La pulsation, le tempo

La question Quelles sont les valeurs du tatum, du tactus et de la mesure ?

Terminologie Le **tatum** est défini comme « la division régulière du temps qui coïncide avec la majorité des débuts de notes » [KEA06]. Le **tactus** est défini comme « la période perceptuellement la plus dominante, [...] la fréquence à laquelle la majorité des gens taperaient du pied ou des mains en phase avec la musique » [Pee07a]. Les indications de tempo éventuellement données par l'auteur se réfèrent généralement au tactus. La **mesure** est définie par le compositeur. La recherche de la mesure correspond à la recherche des temps forts parmi les pulsations détectées (les tactus *a priori*).

Une illustration de ces termes est donnée sur la figure 2.9, qui correspond au début du deuxième petit prélude de Bach : le tatum (en bleu) correspond à la croche, le tactus (en vert) à la noire ; la mesure (en bordeaux) est de 4 noires.

Dans le chiffrage de la mesure, le tactus correspond souvent au dénominateur (4 pour une noire, 8 pour une croche...), alors que la détection des temps forts correspond au numérateur [GD05]. Il est cependant intéressant de noter que la perception du tempo (du tactus) est différente d'une personne à l'autre. L'âge, les connaissances musicales, mais aussi le moment de la journée sont sources de différences [Dra93, DW00, Lap00]. Cependant, ces variations de tempo sont la plupart du temps liées par des rapports deux, un-demi, trois, ou un-tiers.

Cette tâche trouve des applications à la fois dans le domaine de l'indexation musicale, ou de l'analyse automatique de musique, mais également dans des domaines plus « exotiques », tels que la synchronisation des lumières sur une musique (concerts, boites de nuit), ou encore la synchronisation d'un diaporama sur les changements de notes.

Historique Les premiers travaux se sont naturellement penchés sur le problème de la détection de la pulsation dans des cas où la pulsation est régulière. Des restrictions ont parfois été rajoutées, en supposant que le chiffrage de la mesure est connu, ou encore que le rythme est donné par des percussions [GM94]. Plus récemment les travaux se sont abstraits de certaines contraintes, pour traiter des cas de plus en plus généraux, jusqu'à arriver au problème de tempos non réguliers [Pee05b, Pee07a].

Parmi les premiers articles publiés sur le sujet, Brown [Bro91b, Bro93], en contexte monophonique, utilise une méthode par calcul de l'autocorrélation : à chaque instant est attribué un poids. Celui-ci correspond à la durée de la note à l'instant de début des

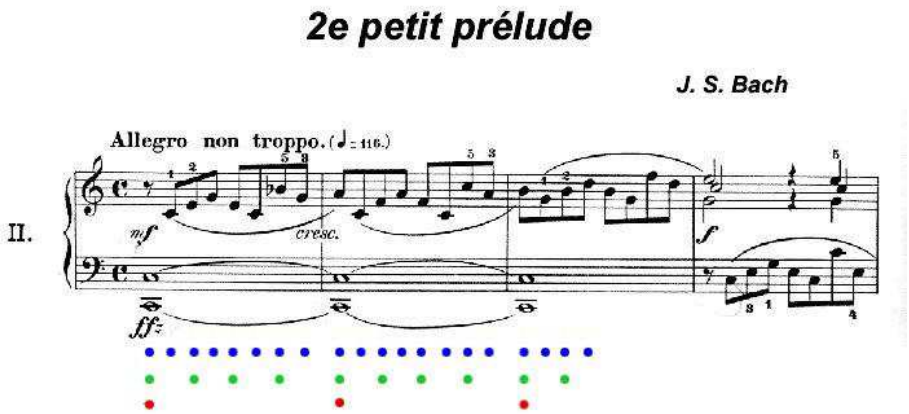


FIG. 2.9 – Exemple des relations entre le *tatum*, en bleu, le *tactus*, en vert, et la *mesure*, en bordeaux, sur le début du deuxième petit prélude de Bach.

notes, 0 ailleurs. En calculant l'autocorrélation de cette représentation, on détermine à la fois la pulsation, qui correspond aux pics de l'autocorrélation, mais aussi la mesure, qui correspond aux maxima.

Méthodes actuelles

Cependant, il est clair que cette méthode ne peut être appliquée que pour des morceaux pour lesquels soit de la transcription (par exemple en MIDI), soit des enregistrements séparés des voix sont disponibles. Cela se ramène alors au cas monophonique.

Pour la détection du rythme par l'analyse de signaux audio polyphoniques, deux approches sont actuellement utilisées : la première, la plus courante, réalise une analyse directe du signal pour extraire le tatum et/ou le tactus et/ou la mesure. La deuxième, plus rare, extrait des informations d'assez haut niveau (les changements d'accords, les débuts de notes, la partition percussive...) et analyse ces informations pour trouver le rythme.

Les différentes étapes de la majorité des algorithmes sont les suivantes :

1. Trouver la bonne représentation du signal (énergie par bandes de fréquences [Sch97], Transformée de Fourier Discrète [Pee05a], représentation temps-fréquence [Kla06]),
2. (*Étape facultative*) Comparer les vecteurs à différents instants (par une fonction d'autocorrélation[Pee05a], de différence [Sch97], de similarité [FU01]...),
3. Analyser fréquentiellement la fonction précédente (bancs de filtres [Sch97], filtres particulaires [HM03], filtres en peigne [KEA06]),
4. En déduire la pulsation.

Pour la deuxième approche [GM99, Got01, ADR04, Pee05b], l'analyse des débuts de notes, changements d'accords et motifs rythmiques est faite à l'aide d'heuristiques, suivie

d'une étape de résolution des ambiguïtés issues des différentes analyses. Cette méthode demande cependant le développement d'outils performants pour l'extraction des « indices ». Les heuristiques sont construites à partir de connaissances musicales relativement bien établies, mais nécessitent une adaptation pour tout nouveau type de musique analysé.

Enfin, des travaux ont été menés, pour estimer conjointement plusieurs paramètres partiellement corrélés : le rythme et les accords [PP08], la tonalité, le rythme et les accords [SW05, Kla03]. Les accords et la tonalité sont naturellement corrélés, puisque le premier contraint les seconds, selon des règles définies pour chaque style de musique. Les accords et le rythme sont également corrélés : les changements d'accords se font principalement sur les temps forts.

Performances Comme précisé précédemment, tout le monde n'a pas la même perception de la pulsation d'un morceau. Pour cette raison, il est généralement admis que les pulsations double (resp. triple) ou moitié (resp. un tiers) de l'annotation sont justes pour les morceaux en binaires (resp. en ternaire).

Pendant la conférence ISMIR 2004, une comparaison de divers algorithmes d'estimation du tempo a été réalisée sur un corpus commun. Ce corpus est décrit dans [GKD+06] et est en partie disponible sur Internet[18]. Il est composé de boucles rythmiques, de danses, et de chansons. Sur ces extraits dont le tempo est fixe, l'accuracy est de l'ordre de 80 % à 90 %. Elle tombe cependant à 50 % à 65 % en considérant faux les doubles, moitié, etc [Pee07a]. Les résultats sont variables selon le type de musique considéré. Le tempo des musiques de danse (rock, tango, valse, cha-cha,...) est en général plus facile à estimer que celui des chansons.

2.5.4 Le genre

La question À quel genre musical appartient cet extrait ?

Terminologie Il est nécessaire de préciser la différence – non évidente – entre deux termes proches couramment utilisés en musique : le « genre » et le « style ». Fabbri [Fab99] définit les deux termes, « genre » et « style » de la façon suivante :

Le genre est « un genre de musique, tel que généralement admis par une communauté pour quelque raison, but ou critère que ce soit, par exemple un ensemble d'événements musicaux dont le cheminement est régi par des règles (de n'importe quelle sorte) acceptées par la communauté ». Le genre est lié à une communauté musicale, qui se reconnaît comme telle au travers de critères propres.

Le style est « un arrangement de paramètres récurrents dans des événements musicaux, qui est typique d'un individu (compositeur, interprète), d'un groupe de musiciens, d'un genre, d'un endroit ou d'une période [historique] ».

[18] http://ismir2004.ismir.net/ISMIR_Contest.html

Il en sort que le genre se réfère plutôt à la construction d'un morceau de musique, alors que le style se réfère plutôt à son interprétation.

De la difficulté de définir le genre La question du genre semble simple dès lors que l'on se limite à rechercher des grandes familles musicales, du type « Rock », « Jazz », « Classique »... La définition des classes est cependant plus compliquée.

Peut se poser la question de la granularité : faut-il distinguer le baroque, la renaissance et le romantique au sein de la musique classique ? Si cette distinction n'est pas faite, le classique représente plusieurs centaines d'années de productions, alors que le Jazz ou le Rock ne représentent que quelques dizaines d'années chacun.

Certains morceaux de musique, surtout dans les musiques actuelles, sont à la frontière entre différents styles et sont donc difficiles à annoter. Le genre devient une notion à la fois ambiguë et subjective [MF06]

Comme le soulignent Lidy *et al.* [LSC+09], cette catégorisation de la musique en genre est très spécifique à la musique occidentale. Pour des musiques africaines, par exemple, la notion de genre a peu d'importance. Il est alors plus pertinent de s'intéresser à la région d'origine, ou à la fonction d'une musique (chant de guerre, naissance, fête, prière, satire, complainte...). McKay et Fujinaga [MF06] s'interrogent ainsi sur la pertinence de la classification en genre. Ils résument les arguments positifs et négatifs ainsi :

Pour

- un certain sens d'un point de vue culturel,
- une classification utile du point de vue de l'utilisateur,
- un vocabulaire relativement accessible pour l'utilisateur lambda.

Contre

- la difficulté d'accord entre les annotateurs, tant sur le choix des catégories que sur l'annotation elle-même,
- une annotation souvent faite pour un artiste ou un album, plutôt que sur chaque morceau de musique,
- le temps nécessaire à l'annotation,
- l'apparition régulière de nouveaux genres, et une évolution de leurs définitions au cours du temps,
- des performances actuelles insuffisantes pour être applicables dans des situations réalistes.

Pachet et Cazaly [PC00] ajoutent à ces inconvénients des problèmes au niveau du vocabulaire.

Malgré tout, les recherches sur ce sujet se poursuivent car, comme le disent Aucoutuier et Pachet [AP03], le genre reste probablement un des descripteurs les plus utilisés par le grand public.

Historique Les premiers travaux semblent dater de la fin des années 1990.

Lambrou *et al.* [LKS+98] utilisent un schéma classique : extraction de paramètres et classification. Les paramètres, calculés sur le signal, sont : des statistiques d'ordre 1 (moyenne, variance, asymétrie et kurtosis), des statistiques d'ordre 2 (second moment angulaire, corrélation et entropie), et le taux de passage à zéro.

Les classifieurs sont de type k-Plus Proches Voisins.

En testant leur système sur 12 extraits appartenant à 3 catégories (Rock, Jazz et Piano), l'accuracy est de l'ordre de 90 %. Certes, le nombre de classes est très faible, et la classe Piano n'est pas comparable avec les deux autres classes.

Soltau *et al.* [SSWW98] cherchent des « séquences d'événements acoustiques ». Chaque séquence est décrite par des paramètres de nature statistique. La classification s'effectue sur ces paramètres statistiques.

L'extraction des événements acoustiques est réalisée par une classification non supervisée en 10 classes, à partir des coefficients cepstraux. Chaque événement (chaque trame d'analyse) est caractérisé par sa classe, et son « activation », qui correspond à son degré d'appartenance à cette classe. Une séquence est ensuite l'enchaînement des activations de ces 10 classes.

Chaque séquence est résumée par les statistiques suivantes : l'occurrence de chaque événement, la co-occurrence de chaque paire d'événements, la tri-occurrence de chaque triplet, la durée de chaque événement, ainsi que la moyenne, le maximum et la variance des activations de chaque événement.

La classification est réalisée par un réseau de neurones.

La base de données est plus consistante que dans l'étude de Lambrou *et al.* : 3 heures de son, réparties en 4 classes (Rock, Pop, Techno et Classique), correspondant à 360 extraits de 30 secondes, également répartis entre les classes. le taux de reconnaissance est de 86 %. Notons que la prise en compte de la dimension temporelle, avec les bi- et tri-grammes, est très importante pour cette tâche.

Méthodes actuelles

Un très large ensemble de paramètres est utilisé actuellement pour ce type de classification. Aucouturier et Pachet [AP03] les classent en trois grandes catégories : les paramètres représentant le timbre, ceux représentant le rythme et ceux construits autour de la fréquence fondamentale.

Les paramètres représentant le timbre sont censés caractériser l'instrumentation. Ce sont :

- les coefficients issus de la Transformée de Fourier [LO03],
- les coefficients cepstraux et les MFCC [BCE+06, ME08, Pee08, Tza08, LOL03],
- les divers paramètres spectraux (Roll-Off, moyenne, variance, asymétrie, kurtosis, centroïde) [BCE+06, Tza08],
- le taux de passage à zéro [BCE+06].

Les paramètres représentant le rythme caractérisent la régularité du rythme, de la pulsation, du tempo,... Ce sont :

- le motif rythmique (« rhythm pattern » en anglais) [LR05, LRP+08], proposé par Rauber et Frühwirth [RF01]. Ce paramètre représente les motifs rythmiques récurrents dans chaque bande spectrale (les 24 bandes de Bark, réparties entre 0 et 15 500 Hz),
- l'histogramme de pulsation [LOL03], qui donne, à l'échelle d'un morceau, l'importance de chaque pulsation possible,
- l'histogramme de rythme [LR05, LRP+08].

Les paramètres construits autour de la fréquence fondamentale décrivent le contenu mélodique et harmonique du signal :

- le chroma vector [Pee08],
- la fréquence fondamentale [LOL03].

D'autres paramètres ont été proposés : une transformée en ondelettes [LOL03], ou encore des paramètres dits « paramètres symboliques » [PdLIn07, LRP+08], calculés sur des données MIDI : le nombre de syncopes (notes ne commençant pas sur un temps), le nombre de notes et leurs durées, la tessiture, des statistiques sur les intervalles présents, des statistiques sur les notes n'appartenant pas à la gamme diatonique.

Notons que, pour de nombreux paramètres calculés sur chaque trame, le but est souvent de les résumer sur le morceau entier, à l'aide d'histogrammes [LOL03, LR05], ou de valeurs statistiques (moyenne, variance, maximum, minimum...) calculées à l'échelle du morceau [PdLIn07, Tza08].

La classification est réalisée avec des outils classiques, tels les SVM [LOL03, LRP+08], les k-PPV [LOL03, PdLIn07], l'approche Bayésienne, avec des GMM [LOL03, Pee08], ou des lois normales [PdLIn07].

Performances La classification en genre est une des tâches pour lesquelles les performances des différents algorithmes sont comparables. En effet, elle a été proposée durant plusieurs années dans la campagne d'évaluation MIREX[19].

Lors de MIREX 2008, cette tâche était évaluée sur deux corpora :

- Une classification sur un corpus « mixte », contenant les genres suivants : Blues, Jazz, Country/Western, Baroque, Classique, Romantique, Électronique, Hip-Hop, Rock et Hard-Rock/Métal.

[19] http://www.music-ir.org/mirex/2009/index.php/Main_Page

– Une classification sur un corpus « latin », contenant les genres suivants : Axé[20], Bachata[21], Boléro[22] Forró[23], Gaúcha[22], Merengue[24], Pagode[25], Sertaneja[26] et Tango[22].

Sur la tâche « mixte », les meilleurs résultats sont de 66,5 % d'accuracy [Tza08]. Il est intéressant de noter que certains genres sont difficilement reconnus, quelle que soit la méthode, tel le Rock (accuracy entre 39 % et 45 %), alors que d'autres sont au contraire toujours bien classés, tel le Hip-Hop (accuracy toujours supérieure à 80 %).

Sur la tâche « latin », les meilleurs résultats sont de 62,7 % d'accuracy [CL08]. Là encore, certains genres semblent extrêmement difficiles à reconnaître (Sertaneja, avec des résultats autour de 20 % d'accuracy, ou Pagode et Forró, avec environ 30 % d'accuracy), alors que d'autres sont très bien classés (Merengue, plus de 80 % d'accuracy pour presque tous les participants).

Ces travaux ont atteint une bonne maturité. Cependant, certains genres restent encore difficiles à reconnaître.

2.5.5 Les émotions dans la musique

La question Quelle(s) émotion(s) cette musique évoque-t-elle ? Le problème est un problème de classification fermé, uni- ou multi-label, selon les cas.

Terminologie En anglais « music mood ». Ce terme, si la communauté s'accorde sur son sens, pose de nombreux problèmes :

– Une même musique n'évoque pas forcément les mêmes émotions chez deux personnes différentes. Pire, elle peut évoquer des émotions différentes chez une même personne, selon le contexte ! Ainsi, une première difficulté apparaît dès l'annotation des corpora. Pour ce faire, traditionnellement plusieurs annotateurs déterminent l'émotion d'un extrait, la décision finale se faisant soit par un vote à la majorité, soit par un vote à l'unanimité. Une alternative est de proposer plusieurs émotions pour un même extrait, permettant ensuite une classification « multi-label ».
– Quel niveau d'émotions retenir comme référence ? Par exemple, faut-il différencier la joie de l'allégresse ? Ainsi, les différents corpora disponibles ne se basent pas tous sur le même nombre d'émotions. Ce nombre peut aller de six [TTKV08] à une vingtaine [CN09]. La question se pose pour la reconnaissance d'émotions tant sur la parole que sur la musique.

[20] Musique populaire – Brésil
[21] Musique romantique – République Dominicaine
[22] Danse
[23] Danse – Brésil
[24] Danse – République Dominicaine
[25] Sous-genre de la Samba
[26] Country – Brésil

– La traduction des termes entre les différentes langues. Par exemple, il est possible de différentier « quiet » et « calm » en anglais, alors que les deux termes se traduisent de la même façon en français.

Des espaces de représentation Différents espaces de représentation des émotions en général ont été proposés. Parmi eux, nombreux sont ceux [Rus80, LD92, Sch99, CN09] qui représentent les émotions dans un espace à deux dimensions. Les deux axes de cet espace représentent la valence et l'activation. La valence mesure la « positivité » de l'émotion (« triste » a une valence plus faible que « joyeux ») ; l'activation mesure la puissance de l'émotion (« en colère » a une excitation plus forte que « déçu », ou encore que « calme »). La figure 2.10 présente un exemple de représentation des émotions dans l'espace valence-activation.

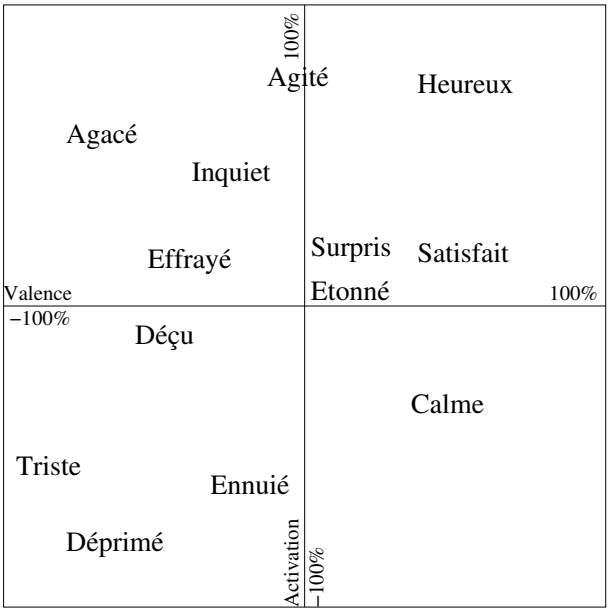


FIG. 2.10 – Un espace de représentation des émotions : l'espace Valence-Activation.

Historique La musique peut créer des émotions chez l'auditeur. La (?) première étude de l'influence émotionnelle de la musique est due à Hevner qui, en 1936 [Hev36], proposait une première représentation des émotions en 8 classes, chaque classe étant définie par une dizaine d'adjectifs de sens proches. Ces classes sont représentées sur un cercle, de façon à ce que deux classes de sens proche soient également proches. Remarquons que cette présentation, comme d'autres plus tard, laisse entendre qu'il y a une « circularité » dans les différentes émotions.

Les premiers travaux sur l'identification automatique des émotions en musique datent de la fin des années 1990, avec un véritable essor de ce sujet au milieu des années 2000.

Méthodes actuelles

Il est intéressant de noter que les algorithmes utilisés pour la classification en émotions sont souvent très proches de ceux utilisés pour la classification en genre. Peeters [Pee08], Mandel et Ellis [ME08], ou encore Tzanetakis [Tza08] ont même proposé d'utiliser le même algorithme pour les deux tâches dans la campagne d'évaluation MIREX 2008.

Ainsi, les paramètres extraits du signal tendent à modéliser le timbre [LO03, TTKV08, Pee08, YLSC08, TTK05] (MFCC et dérivées, centroïde spectral, Roll-Off, flux spectral...), le rythme [TTKV08, LO03, LLZ06, YLSC08] (motifs rythmiques, histogramme de pulsation, sa régularité, sa valeur et son intensité...), l'intensité [LLZ06, YLSC08] (énergie à court terme, et énergie dans chaque sous-bande spectrale), les fréquences présentes [YLSC08] (les accords, le nombre de fréquences fondamentales, les dissonances...). Yang *et al.* [YLSC08] utilisent également la transformée en ondelettes.

Tout comme pour la classification en genre, les paramètres calculés sur chaque trame (les paramètres spectraux notamment) sont souvent résumés à l'échelle du morceau par leur moyenne et leur variance [LLZ06, YLSC08]. Trohidis *et al.* [TTKV08] y ajoutent la moyenne de la variance et la variance de la variance.

Dans les études mono-label, les outils de classification classiques sont utilisés : GMM ou SVM. Certaines études utilisent une classification hiérarchique, la hiérarchie étant souvent décidée en fonction de représentations usuelles des émotions en grandes classes : par exemple, émotions positives/négatives, puis émotions fortes/faibles, etc.

Dans l'étude multi-label menée par Trohidis *et al.* [TTKV08], la classification est réalisée en utilisant des outils usuels pour ces problèmes. Une possibilité est, par exemple, de considérer le problème comme une succession de problèmes binaires, indépendants les uns des autres. La plupart des outils de classification binaire peuvent être étendus au cas multi-label, comme par exemple les k-Plus Proches Voisins ou les SVM.

Enfin, notons une étude intéressante de Hu et Downie [HJ07], qui vise à analyser les éventuelles corrélations entre genre et émotion. Certes, l'émotion la plus souvent associée à un genre est souvent pertinente (par exemple les paires « Électronique-Hypnotique », ou « Gospel-Spirituel »), mais il s'avère que chaque genre est associé de manière significative à une dizaine d'émotions en moyenne. Il en ressort que les relations entre genre et émotion sont trop variables pour être utilisées en classification de genre, mais pourraient cependant être utilisées pour la classification en émotion. Ils concluent également que de nombreuses émotions sont trop proches pour être correctement distinguées, que ce soit par les algorithmes ou lors de l'annotation. Ainsi, ils proposent de n'utiliser que quelques (cinq ou six) groupes d'émotions comme labels.

Performances En reconnaissance d'émotions à partir de la parole, les corpora utilisés, comme le corpus de Berlin[27] [BPR+05], sont composés de phrases dites par des acteurs,

[27] accessible en ligne à l'adresse suivante : http://www.expressive-speech.net/emodb/

qui simulent différentes émotions. Il est très difficile d'obtenir des données « naturelles ». Dans le cas de la musique, ce problème est moins saillant, puisque de très nombreux enregistrements sont disponibles.

Différents corpora ont été étiquetés pour cette tâche. Trohidis *et al.* [TTKV08] ont annoté 593 chansons en 6 émotions : étonnement-surprise, heureux-satisfait, relaxant-calme, tranquille, triste-solitaire, et peur-colère. Ces extraits sont de différents styles : Classique, Reggae, Rock, Pop, Hip-Hop, Techno et Jazz. Les labels ont été attribués par 3 experts, issus d'une école de musique ; seuls les extraits annotés à l'unanimité ont été retenus dans le corpus final, qui est accessible en ligne[28].

Trohidis *et al.* obtiennent une accuracy de 73 à 87 %, selon l'émotion qu'ils considèrent. Il s'avère que dans leur étude, l'émotion la plus facile à trouver est relaxant-calme, alors que la plus difficile est heureux-satisfait.

La campagne d'évaluation MIREX a introduit cette tâche en 2007. Lors de la campagne 2008, cinq classes ont été construites selon les recommandations de Hu et Downie [HJ07] :

Groupe 1 : passionné, exaltant, confiant, tapageur, bruyant/turbulent,

Groupe 2 : exubérant, joyeux, amusant, doux, aimable,

Groupe 3 : recherché, poignant, mélancolique, doux-amer, automnal, maussade,

Groupe 4 : humoristique, stupide, excentrique, saugrenu, spirituel, narquois,

Groupe 5 : agressif, fougueux, tendu, intense, explosif, viscéral.

Les meilleurs résultats sont de 63,7 % d'accuracy [Pee08]. Comme pour la classification en genre, il existe une très grande disparité entre classes : l'accuracy est autour de 50 % pour les groupes 1, 2 et 4, alors qu'elle monte à plus de 80 %, pour les groupes 3 et 5.

2.5.6 L'identité du chanteur

La question « Lequel, parmi ces N artistes connus dans notre base de données, est en train de chanter ? » Il s'agit d'un problème fermé, les artistes sont connus à l'avance, on dispose la plupart du temps d'exemples de leur voix, que ce soient des exemples solos ou accompagnés.

Historique - Analyse des capacités humaines Les premiers travaux, notamment menés par Cleveland [Cle77], ou encore Bloothooft *et al.* [BR88] se sont attachés à caractériser le timbre de chaque voix. Similairement à la définition du timbre des instruments, le timbre d'une voix est défini comme « la caractéristique de la voix qui distingue une voix d'une autre quand la hauteur de la note et la voyelle chantée sont les mêmes ». Ces études se basent sur l'hypothèse implicite que chaque voix est caractérisée par un timbre unique, qui est fonction des caractéristiques physiologiques de la personne.

[28] http://mlkd.csd.auth.gr/multilabel.html

Cependant, une étude de 2001, menée par Erickson *et al.* [EPH01] s'intéresse à la capacité humaine de distinguer les voix (et donc les timbres), en fonction de la hauteur de la note chantée. Il n'est pas évident *a priori* que deux personnes dont les voix sont perçues comme différentes pour une certaine note, ne seront pas confondues pour une autre note. Les auteurs mènent l'expérience suivante : présentons à un auditeur trois notes (A, B et C). Deux d'entre elles (A et B) sont chantées par la même personne, la troisième (C) est chantée par quelqu'un d'autre. L'auditeur doit retrouver l'intrus. Il s'avère que plus les notes A et B sont éloignées en terme de hauteur, et plus les performances de l'auditeur sont faibles. Si les notes A et B sont éloignées de deux tons (Sol4 et Si4), le taux de réussite est presque de 100 %. Par contre, si les notes sont éloignées de plus d'une octaves (Do4 et Fa5, ou encore La3 et La5), le taux de réussite tombe en dessous de 50 %, voire même en dessous de 33 % (valeur qui correspond au hasard). Il semblerait que le timbre soit variable en fonction de la hauteur de la note chantée.

Enfin, pour conclure cet historique, nous notons, comme le font très justement remarquer Nwe *et al.* [NL07b] que le problème de l'identification du chanteur dans les musiques commerciales[29] est compliqué par le fait que le chant est très souvent accompagné par des instruments de musique.

Méthodes actuelles

Les recherches peuvent schématiquement être séparées en deux catégories distinctes : les études sur le chant solo, et celles portant sur le chant accompagné, contexte certes plus réaliste, mais aussi *a priori* nettement plus difficile !

Dans tous les travaux concernant le chant accompagné, le schéma général pour la phase de reconnaissance est le suivant :

1. Extraction des zones de chant (l'état de l'art pour cette tâche est présenté dans la partie 4.2). Il est à noter cependant que dans ce cadre, trouver toutes les zones de chant n'est pas forcément indispensable. La recherche de précision est privilégiée sur celle du rappel pour évaluer la pertinence de l'algorithme mis en œuvre dans cette première étape.
2. (*Étape facultative*) Séparation de sources.
3. Extraction de paramètres sur les zones de chant.
4. Classification.

Les travaux s'intéressant au chant solo excluent naturellement les deux premières étapes de leurs algorithmes.

S'inspirant des nombreux travaux réalisés en identification du locuteur, Zhang [Zha03] applique un schéma classique : le signal est caractérisé par les paramètres « phares » du

[29] musiques non enregistrées pour les besoin de la recherche

traitement de la parole, les MFCC, ainsi que par des LPC. Berenzweig *et al.* [BEL02] utilisent également comme paramétrisation les MFCC, tout en admettant qu'il y a sûrement de meilleurs paramètres à trouver, puisqu'il s'agit ici de chant et non de parole. Kim et Whitman[KW02] paramétrisent le signal avec des LPC ou des « Wraped Linear Prediction models ».

D'autres paramètres ont depuis été étudiés : le vibrato (voir partie 4.3.1) avec sa fréquence, sa stabilité, sa régularité, par Nwe et Li [NL07b], l'enveloppe spectrale [BW04] et la « Composite Transfert Function » (les fréquences et amplitudes instantanées des partiels présents dans le signal) par Bartsch et Wakefield [WB03].

Il y a naturellement, et ce dans tous les algorithmes que nous avons recensés dans la littérature, une phase d'apprentissage de modèles pour chacun des chanteurs. Cet apprentissage peut se faire sur des zones de chant sélectionnées manuellement [Zha03], ou sur des extraits musicaux entiers [KW02].

Pour la modélisation de chaque chanteur, les outils les plus utilisés sont les GMM [Zha03, KW02] et les SVM [KW02], mais Nwe et Li [NL07b] proposent par exemple d'utiliser des HMM, qui permettent de prendre en compte l'évolution temporelle des paramètres.

Enfin, Maddage *et al.* [MXW04] proposent, de façon très pertinente, d'améliorer leur système en tenant compte, pour l'étape de regroupement, non seulement des similarités de la voix, mais également des similarités de style. En effet, les chanteurs (ou groupes de musique) ont souvent un style musical « à eux ». Berenzweig *et al.* évoquent [BEL02] d'ailleurs l'« effet album » : comme il y a une certaine homogénéité musicale (instruments, styles...) au sein d'un album, il faut prendre garde à ce que le classifieur reconnaisse bien le chanteur, et non l'album !

Performances Dans le cas du chant solo, le corpus est « fait maison », les enregistrements sont faits en studio. Cela offre la possibilité d'avoir la même mélodie (ou la même phrase) chantée par plusieurs personnes [WB03]. Dans le cas du chant accompagné, le corpus est composé d'extraits commerciaux.

Les performances actuelles sont de l'ordre de 15 à 20 % d'erreur, pour l'identification de chanteurs solo. Pour l'identification du chanteur accompagné, Nwe et Li [NL07b] atteignent 16 % d'erreur. Ces résultats peuvent paraître surprenants, puisqu'il semble aussi facile d'identifier un chanteur solo qu'accompagné. Il faut cependant se rappeler que les corpora ne sont pas du tout les mêmes : dans le cas de chanteurs solo, un certain nombre de personne chantent les mêmes extraits, alors que pour le chant accompagné, il s'agit d'extraits commerciaux. Ainsi, l'« effet album » peut exister dans le deuxième cas, mais pas dans le premier.

Vers l'identification de plusieurs chanteurs Nous tenons à citer ici une dernière étude, réalisée par Tsai *et al.* [TLL08], qui cherche à identifier deux chanteurs intervenant

au cours d'un même morceau (simultanément ou non). La musique contenant des paroles ne se résume pas à des morceaux contenant un seul chanteur, il peut y avoir plusieurs voix, qu'elles soient de même importance ou que l'une soit prédominante. Pour simplifier le problème, les auteurs s'intéressent à des duos *a capella*. Après avoir déterminé si l'extrait est à une ou deux voix, ils cherchent le meilleur modèle pour représenter l'extrait. Dans le cas de chant solo, ils possèdent un modèle par chanteur. Dans le cas des duos, ils ont besoin d'un modèle pour chacune des paires possibles. Comme il n'est pas toujours possible de réunir dans la base d'apprentissage des exemples de chant *a capella* pour toutes les paires, ils cherchent à créer artificiellement ces modèles, à partir des données dont ils disposent pour chacun des chanteurs *a capella*.

Les tests sont effectués sur une base de données contenant des duos vocaux non accompagnés. Lorsque les mélodies et paroles de test sont toutes différentes, de même que celles de l'apprentissage, 71 % des chanteurs, et 43 % des paires de chanteurs sont correctement identifiés.

2.5.7 Les transcriptions – la mélodie

Les travaux sur ce sujet se divisent en deux grandes catégories : la transcription de la partie percussive, et la transcription d'une ligne mélodique dominante. La ligne mélodique peut être un instrument, ou une voix chantée.

2.5.7.1 La partie percussive

La question Il s'agit de fournir une partition de la partie jouée par les percussions.

L'extraction de la partie percussive a de nombreuses applications : elle peut permettre l'identification du genre [DGW04] (le jazz et le rock sont par exemple très différents de ce point de vue), la recherche par le rythme d'extrait musicaux dans une base de données [KBT04, GR05], ou encore servir d'outil pour la composition ou le mixage de musiques [PK03, GR05].

Historique Les recherches se sont tout d'abord attelées à une version simplifiée du problème : la transcription d'une partie percussive solo. Par la suite, le problème s'est complexifié, quand les chercheurs se sont penchés sur le problème de la transcription de la partie percussive en contexte polyphonique. La très grande majorité des recherches se sont concentrées sur quelques types de percussions, à savoir celles rencontrées dans les musiques pop et rock, autrement dit la batterie : cela inclut la caisse claire, la grosse caisse, les toms, le charleston et les cymbales. Les différences entre ces instruments sont à la fois d'ordre fréquentiel (la caisse claire est plus aiguë que les toms, eux mêmes plus aigus que la grosse caisse), et d'ordre temporel (durée du son produit) [HYG02].

Méthodes actuelles

Comme Gillet et Richard le résument dans leur article [GR08], trois approches sont actuellement utilisées pour cette tâche :

- « segmentation puis classification »,
- « recherche de motif et adaptation »,
- « séparation puis détection ».

Dans la première approche, l'étape de segmentation du signal audio en « événements », précède l'étape de classification dans laquelle leur sont attribué un instrument. La segmentation est une détection soit des débuts de notes (tâche difficile), soit du tatum (voir partie 2.5.3, tâche plus facile), mais une erreur de sur-segmentation peut alors être très gênante. L'étape de classification commence par l'extraction de paramètres. La plupart des paramètres « classiques » ont été testés : les MFCC, leurs dérivées première et seconde, ainsi que les moyenne et variance de chacun des coefficients [GR04, PK03], l'énergie par bandes de fréquence [GR04, HDG03], les quatre premiers moments du spectre [PR02]... Le module de décision reprend également les outils classiques : en mode supervisé ou non, avec des GMM, des SVM, ou des arbres de décision. Cette méthode générale (segmentation et classification) a été développée au départ pour la transcription de percussion solo, et s'avère effectivement très efficace dans ce cas [GR04]. Dans le cas de musique polyphonique, cette approche est nettement plus difficile à mettre en œuvre [TDDB05], puisque les autres sources vont non seulement rendre plus difficile la segmentation, mais également influer sur les paramètres utilisés pour la classification.

La seconde approche, nécessite de disposer d'un exemple de chaque instrument pour en déduire un motif caractéristique. Ce motif, temporel [ZPDG02] ou fréquentiel [YGO04], est recherché dans le signal. Une fois qu'il est trouvé, on peut l'adapter, chercher le nouveau motif, le réadapter, et ainsi de suite. Notons que cette méthode sous-entend que le motif de chaque instrument reste toujours le même.

La troisième approche, fait appel aux méthodes de séparation de sources pour extraire la partie de percussion, puis la transcrire. Si on dispose d'autant de canaux que de sources, alors des méthodes telles que l'Analyse en Composantes Indépendantes (ICA) peuvent être utilisées. Cependant, dans la majorité des cas, on ne dispose que de deux signaux (enregistrements stéréos), voire même que d'un seul signal (enregistrements monos). La plupart des algorithmes se sont concentrés sur ce dernier cas. Après avoir séparé les différentes sources, et avoir identifié celles correspondant aux percussions, il s'agit idéalement d'une simple tâche de transcription monophonique. Pour plus de détails sur les méthodes de séparation de sources, on pourra se référer à la synthèse de Virtanen, dans le chapitre 9 de l'ouvrage dirigé par Klapuri et Davy [KD06], ou encore aux campagnes d'évaluation SiSec [VAB09].

Quelques travaux ont tenté de modéliser les relations entre les rythmes des différentes percussions, en apportant ainsi une connaissance musicale à leur algorithme. Cela s'est fait par exemple par la prise en compte du contexte à court terme dans un HMM [GR05], par la recherche de motifs rythmiques répétés (typiquement la pulsation, donnée par le charleston) [PK03], ou encore par l'apport de connaissances musicologiques (la manière dont le morceau a été composé) [PK03].

Performances Les scores sont actuellement de l'ordre de 65 à 80 % (selon l'instrument considéré) pour l'accuracy de la transcription, dans des extraits de musique polyphonique.

2.5.7.2 La mélodie principale

La question Il s'agit d'obtenir une transcription de la mélodie principale.

Une application de cet outil est la recherche d'extraits musicaux en fredonnant la mélodie, « Query by Humming » en anglais. Il est en effet nécessaire de savoir extraire la mélodie principale d'un extrait pour pouvoir la reconnaître automatiquement.

Historique Un des précurseurs en la matière fut Goto [Got99, Got04], qui développa en 1999 sa méthode « PreFEst », qui permet d'estimer la fréquence fondamentale prédominante dans de la musique polyphonique et d'obtenir la ligne mélodique principale, ainsi que la ligne de basse. L'idée est de détecter les fréquences présentes[30], puis de rechercher la ligne mélodique caractérisée par des intensités fortes, et une certaine continuité temporelle. Pour distinguer la mélodie de la ligne de basse, l'auteur propose simplement de séparer les hautes et moyennes fréquences (mélodies) des basses fréquences (ligne de basse).

Méthodes actuelles

Deux approches sont possibles :

- détecter les fréquences présentes, puis les assigner au bon instrument (ou à la bonne voix),
- faire de la séparation de sources, puis transcrire chacune des sources monophoniques.

Les campagnes d'évaluation MIREX 2004 et 2005, ont mis en évidence de nombreux systèmes (14 au total, avec plusieurs systèmes pour certains participants) : Dressler [Dre05], Marolt [Mar04b], Goto [Got04], Ryynänen *et al.* [RK05], Poliner *et al.* [PE05], Paiva *et al.* [PMC04], et Vincent *et al.* [VP05]. Une analyse des résultats de ces deux campagnes est proposée dans l'article de Poliner *et al.* [PEE+07b]

Les algorithmes proposés lors de ces campagnes suivent tous l'approche directe : détecter les fréquences présentes à l'aide d'un algorithme de « multipitch », puis sélectionner

[30] Pour ce qui est de la détection des fréquences présentes (polyphoniques ou monophoniques), on se référera par exemple à la thèse de Yeh [Yeh08].

la bonne fréquence (si elle existe, c'est-à-dire s'il y a une mélodie principale). Les algorithmes « multipitch » se basent globalement : soit sur la Transformée de Fourier à Court Terme [Dre05, Mar04b, Got04, RK05, PE05], soit sur le corrélogramme [PMC04], soit sur l'algorithme YIN (décrit dans la partie 3.3.1 [VP05]). Ces paramètres sont analysés, pour aboutir à la détection de 1 à 5 fréquences fondamentales. Des étapes de détection des débuts de notes, et de suivi sont éventuellement rajoutées pour arriver au résultat final.

Plus récemment, Durrieu *et al.* [DOF+09] ont utilisé l'approche type « séparation de sources ». Dans ce cas, les données sont stéréophoniques. Pour plus de détails, on se référera à l'ouvrage de Klapuri et Davy [KD06] pour les méthodes, et à l'article de Vincent *et al.* [VAB09] pour la campagne d'évaluation SiSec 2008[31].

Performances Les corpora de test utilisés lors des campagnes MIREX, contiennent divers genres musicaux (Pop, Jazz, Classique, R&B, Rock), la mélodie principale étant tenue par des instruments divers : voix (hommes, femmes et synthétiques), saxophone, guitare et instruments synthétiques.

L'évaluation est faite sur les valeurs estimées de fréquences fondamentales, à 25 *cents* près (soit un huitième de ton) – la mélodie s'en déduisant en arrondissant les valeurs de fréquences fondamentales aux notes les plus proches. Les meilleurs résultats sont obtenus par Dressler avec 71,4 % d'accuracy globale, et Ryynänen *et al.* avec 74,1 % d'accuracy en ramenant les notes sur une seule octave.

2.5.7.3 Le cas particulier du chant

La question L'extraction de la mélodie chantée est un cas particulier de l'extraction de la mélodie. Cependant, nous choisissons d'y consacrer un paragraphe pour les raisons suivantes : le chant est souvent la mélodie que l'auditeur retient, et la voix humaine chantée présente des caractéristiques particulières qui ont mené au développement de méthodes spécifiques à cette tâche.

Les systèmes de « Query by Humming » se sont notamment concentrés sur ce problème, qui diffère, même pour des extraits monophoniques, de la transcription des instruments. La voix chantée est moins stable que les notes produites par des instruments, soit que la personne chante faux, soit qu'elle change brusquement de ton, soit tout simplement parce qu'elle contient par nature du vibrato. Pour un état de l'art sur la transcription du chant solo, on pourra se référer au chapitre 12, écrit par Ryynänen de l'ouvrage de Klapuri et Davy [KD06].

Méthodes actuelles

Pour l'extraction de la mélodie chantée en contexte polyphonique, les systèmes proposés

[31] http://sisec.wiki.irisa.fr/tiki-index.php

sont soit des systèmes dédiés à cette tâche [GB08], soit des adaptations de systèmes d'extraction de la mélodie principale [RK06, DRD08, DOF+09, FKG+06].

Ryynänen et Klapuri [RK06] contraignent leurs modèles à la seule étendue de la voix (deux octaves), et apprennent les modèles acoustiques et les modèles de transitions entre notes sur un corpus approprié (RWC Popular Music Database[32]).

Durrieu *et al.* modifient leur méthode de séparation de sources : le filtre utilisé pour détecter la voix prend en compte le fait que celle-ci comporte des formants [DRD08] dans le domaine spectral, ou encore des parties non voisées [DOF+09].

Fujihara *et al.* [FKG+06] se basent sur l'estimateur PreFEst [Got99, Got04], présenté précédemment. Une approche probabiliste fondée sur des GMM leur permet de mesurer la probabilité, pour chaque fréquence fondamentale, qu'elle ait été produite par une voix. Un algorithme de suivi, basé sur la proximité de la trajectoire, de la vraisemblance de la fréquence fondamentale et de la probabilité que celle-ci appartienne à une voix donne la ligne mélodique finale.

Gómez et Bonada [GB08], s'intéressent à la transcription du flamenco et ont développé une méthode dédiée, qui prend en compte les particularités de la voix dans le flamenco : entre autres, la présence d'air dans la voix, l'instabilité de la fréquence fondamentale, l'étendue restreinte de la voix (une sixte !), ou encore l'utilisation d'intervalles plus petits que ceux habituellement utilisés dans la musique occidentale. La méthode est ensuite classique, avec l'extraction de paramètres qui permettent l'estimation de la fréquence fondamentale, et la transcription en notes. Les auteurs y ajoutent une estimation du vibrato.

Performances Durrieu *et al.* testent leur algorithme sur les données MIREX 2004 contenant du chant. Ils obtiennent une accuracy globale de 70 %, tout à fait comparable aux résultats obtenus par les participants à la campagne 2006 sur les mêmes données. Pour la transcription du chant monophonique, les résultats sont de l'ordre de 13 % d'erreur [KD06].

2.5.8 Les transcriptions – la partition

La question Le problème est compliqué, puisqu'il s'agit de transcrire toute la partition, c'est-à-dire à la fois la mélodie principale, mais aussi l'accompagnement. Il peut également être vu comme une prolongation de la tâche d'estimation de fréquences fondamentales multiples : il s'agit en plus de les suivre et de les assigner au bon instrument (ou à la bonne voix).

[32] Real World Computing Music Database, disponible en ligne : http://staff.aist.go.jp/m.goto/RWC-MDB/

Historique Une des premières études a été menée par Moorer [Moo77], pour la transcription de duos. Emiya *et al.* [EBD08], Raphael [Rap02], Poliner *et al.* [PE07], Marolt [Mar04a] et Bello *et al.* [BDS06] se sont intéressés à la transcription des partitions pour piano. Ce problème particulier a été étudié pour plusieurs raisons : des données sont faciles à trouver, tant au niveau monophonique que polyphonique ; la polyphonie est y particulièrement présente, puisque de nombreuses notes peuvent être jouées simultanément.

Méthodes actuelles
Les méthodes utilisées suivent le même schéma que pour la transcription de la mélodie principale : recherche et suivi des fréquences présentes, puis transcription en notes.

La recherche des fréquences se fait la plupart du temps dans le domaine fréquentiel. Cependant, certains auteurs proposent d'y ajouter d'autres paramètres : issus du domaine temporel pour Bello *et al.* [BDS06], ou un modèle de perception auditive pour Marolt [Mar04a]. Pour le suivi, Raphael [Rap02], Emiya *et al.* [EBD08], et Poliner *et al.* [PE07] proposent d'utiliser des Modèles de Markov Cachés. Pour Emiya et Poliner, il s'agit de modéliser les durées des notes, et les transitions entre notes. Raphael se place à un plus haut niveau, puisqu'il s'agit pour lui de modéliser les transitions entre accords.

La transcription en notes peut sembler évidente au premier abord, puisqu'il s'agit simplement de déterminer à quelle note chacune des fréquences détectées correspond. Cependant, le cas des notes répétées pose problème. Il faut en effet les considérer comme deux notes différentes, et bien repérer l'instant de césure. Marolt, ou Emiya *et al.* ont proposé des solutions à ce problème. Pour Emiya *et al.*, ceci est réalisé en analysant la puissance de la note : si sa puissance varie de plus que 3 dB (seuil déterminé empiriquement), ils considèrent qu'il y a deux notes successives. Marolt a envisagé une solution proche : analyser l'amplitude du premier harmonique de la note. Cependant, les recouvrements dans le domaine spectral l'ont mené à utiliser un perceptron multicouche, qui analyse les changements d'amplitudes.

Performances Pour l'évaluation de cette tâche, il est possible d'envisager deux approches. On peut se placer du point de vue de la détection multipitch, et analyser, pour chaque trame (100 ms par exemple), quelles notes ont été trouvées. On peut au contraire se placer du point de vue de l'interprète, et analyser si les débuts de note ont correctement été trouvés. Les critères de performances sont différents selon les auteurs : la F-mesure, le nombre de « Vrai positifs », de « Faux positifs », l'accuracy...

Les évaluations sont menées sur deux grands types de données : des fichiers MIDI, et des enregistrements commerciaux (ou de qualité commerciale). L'avantage des fichiers MIDI est que l'annotation est immédiate ; l'avantage des enregistrements commerciaux est la conformité avec la réalité et leur grande variété. Il est à noter une base de données

de piano MIDI réalisée par B. Krueger, accessible en ligne[33]. On remarquera que les évaluations sont menées sur de la musique classique : Mozart, Chopin, Beethoven, Bach, Debussy, Ravel sont parmi les compositeurs souvent analysés.

Les résultats obtenus par les différents systèmes sont de l'ordre de 65 % pour la F-mesure sur des enregistrements réels. En terme de Vrai positifs (TP), Faux positifs (FP), les meilleurs résultats sont : $TP \simeq 80$ % et $FP \simeq 12 - 15$ %. Emiya *et al.* remarquent cependant que leurs performances sont nettement meilleures pour la musique lente ou avec peu de notes simultanées : elles peuvent atteindre voire même dépasser 90 % pour la F-mesure !

2.5.9 Les transcriptions – la suite d'accords

La question Comment obtenir une transcription de la suite d'accords ?

Terminologie - De la difficulté de se mettre d'accord (!) sur l'annotation. La définition des accords est difficile à fournir :

- Une note tenue fait-elle partie de deux accords ?
- À partir de combien de nouvelles notes un nouvel accord est-il créé ? Ces deux questions s'interrogent sur la position des accords, et sur leur définition. Une solution pour déterminer la position des accords peut être de considérer qu'ils sont aux débuts de chaque mesure, ou à chaque temps (tactus).
- Combien de notes y a-t-il par accord : trois, quatre ? Cette question n'est pas aussi anecdotique qu'il peut paraître, puisque considérer uniquement trois notes implique qu'on exclut les accords de septième, alors qu'ils sont possibles si on considère qu'il peut y avoir jusqu'à quatre notes par accord.
- La méthode de notation des accords pose également problème. Harte *et al.* [HSAG05] abordent la question de façon poussée dans leur article : faut-il noter tous les intervalles, ou certains (comme la tierce par exemple) peuvent-ils être sous-entendus, comme dans les chiffrages d'accords faits sur partition pour les musiciens ? Faut-il noter la note de base de l'accord selon sa valeur absolue (Do, par exemple), ou selon sa valeur relative dans la gamme (tonique, ou dominante, par exemple) ? La figure 2.11[34] résume les diverses possibilités d'annotation que Harte *et al.* [HSAG05] ont envisagées dans leur étude.

Ces difficultés rencontrées au niveau de la définition d'un accord se répercutent sur la production de bases de données annotées. Non seulement elle est coûteuse, mais elle n'est de plus pas toujours consensuelle.

[33] La Classical Piano MIDI Page est accessible à l'adresse suivante : http://www.piano-midi.de/
[34] Figure extraite de l'article de Harte *et al.* [HSAG05]

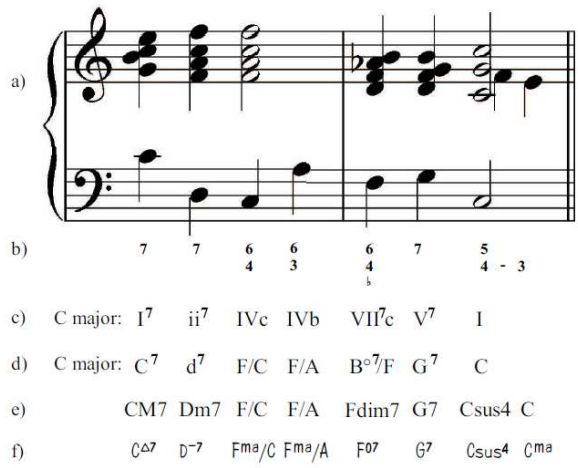


FIG. 2.11 – Différentes possibilités de notation des accords. a) Partition, b) Basse chiffrée, c) Notation romane, d) Notation anglosaxone, e) Notation « guitare », f) Notation « Jazz ».

Méthodes actuelles

Tout comme pour l'identification de la tonalité, les paramètres utilisés ici tendent à refléter les notes de la gamme. Il s'agit par exemple du chroma vector [SE03, PP07, LS08] (voir partie 2.2.2). Nawab *et al.* [NAW01], dans un des premiers travaux réalisés sur le sujet, proposent d'utiliser le Constant-Q Spectrum [Bro91a] (voir 2.2.1). Enfin, des tests ont été réalisés avec des paramètres plus classiques : les MFCC par She et Ellis [SE03], qui ont montré que ces paramètres sont moins efficaces que le chroma vector, et la transformée en ondelettes par Su et Jeng [SJ01].

Harte *et al.* [HSG06] ont proposé un nouveau paramètre : le « tonal centroïde vector », ou centroïde tonal (voir partie 2.2.2). En mesurant la distance euclidienne entre deux centroïdes successifs, ils détectent ainsi les changements harmoniques, c'est-à-dire les changements d'accords. Lee et Slaney [LS08] ont par la suite utilisé ce paramètre pour la transcription des accords : en plus de détecter les changements d'accords, ils ont montré que ce nouveau paramètre est également efficace pour la transcription d'accords, et surpasse le traditionnel chroma vector.

Au niveau des méthodes de reconnaissance, de nombreux travaux introduisent la dimension temporelle pour modéliser les relations entre accords. Des HMM sont appris pour prendre en compte les transitions possibles entre accords [LS08, PP07, LS06, SE03]. Ainsi, on n'identifie pas *un accord* solitaire, mais *une suite d'accords*.

Mauch *et al.* [MDH+07] ont en parallèle commencé à construire un « dictionnaire » des enchaînements d'accords, pour découvrir des séquences d'accords caractéristiques (des « chord idioms », qu'on pourrait traduire par « accords idiomatiques », ou « accords typiques ») d'un style de musique. Ils ont jusqu'à présent étudié deux styles, à partir des

transcriptions d'accords : le Rock, via le corpus des Beatles [HSAG05], et le Jazz, via un des Real Books [Var04][35]. Notons cependant que l'utilisation de tels dictionnaires implique une connaissance *a priori* du genre de musique, soit par le biais d'un pré-traitement qui détermine le genre, soit par intérêt pour une collection musicale particulière.

D'autres études prennent en compte d'autres aspects de la musique. Papadopoulos et Peeters [PP08] utilisent le fait que les changements d'accords se font souvent sur le temps ; il prédisent donc conjointement la pulsation et la suite d'accords. Lee et Slaney [LS08] estiment conjointement la tonalité et la suite d'accords. Ceci leur permet d'adapter les modèles de reconnaissance d'accords à chaque tonalité.

Enfin, Zhang et Gerhard [ZG08] ont proposé d'introduire des connaissance *a priori* sur la composition de l'orchestre, de façon à pouvoir contraindre les notes jouées. Par exemple, si on a un ensemble vocal traditionnel, chacun des chanteurs (soprano, alto, ténor et basse) a un registre limité. De plus, la voix de basse jouera ou chantera souvent la note fondamentale de l'accord.

Performances Les métriques d'évaluation peuvent, elles aussi, engendrer de nombreuses discussions. La solution la plus simple est de considérer qu'un accord détecté est vrai ou faux. Une solution plus souple est de donner un poids aux différentes erreurs. Par exemple : est-il plus grave de confondre Do Majeur et Do mineur, ou Do Majeur et sa relative La mineur ?

Un corpus fréquemment utilisé consiste en 180 chansons des Beatles (leurs 12 albums studio), annotées en accords. Ce corpus a été développé à l'université du Queen Mary par Harte *et al.* [HSAG05]. Les performances obtenues sur ce corpus sont de l'ordre de 70 % à 80,5 % d'accuracy [LS08, PP07].

Papadopoulos et Peeters [PP07] remarquent que l'écart type des performances est très grand : les performances peuvent être très mauvaises sur des morceaux dans lesquels l'harmonisation n'est pas classique : par exemple s'il y a beaucoup d'accords dans lesquels il manque une note (la tierce), il est alors difficile de déterminer s'ils sont majeurs ou mineurs. Ce grand écart type est confirmé par les résultats de Lee et Slaney [LS08], qui passent de 69 % à 86 %, selon qu'ils testent leur algorithme sur l'un ou l'autre CD du corpus des Beatles.

Actuellement, les résultats donnés par les divers algorithmes ne précisent pas les renversements : on ne précise pas l'ordre des notes.

[35] Les Real Books originaux sont des transcriptions – illégales – de nombreux standards de jazz, réalisées par des étudiants du Berklee College of Music dans les années 1970. Les New Real Books en sont une version plus récente – et légale.

2.5.10 Les transcriptions : les paroles

La question Nombre de recherches de musique se font par le biais des paroles d'une chanson. Il est donc nécessaire d'en obtenir une transcription.

Historique La transcription des paroles n'en est actuellement qu'à ses débuts. Les chercheurs aimeraient s'inspirer des travaux déjà réalisés en transcription de la parole non chantée. Cependant, plusieurs problèmes surgissent, et nécessitent de grosses modifications et adaptations des outils : la fréquence fondamentale change beaucoup, les durées des sons s'adaptent au rythme de la musique et sont donc modifiées, les phrases sont de la poésie et ne respectent donc pas toujours les règles de grammaire, plusieurs langues peuvent être mélangées...

Ces divers problèmes expliquent sûrement la rareté des recherches publiées sur le sujet. Les premiers travaux sont d'ailleurs restreints à l'alignement du texte sur la musique [LCB99, FGO+06, MV08].

Méthodes actuelles
Parmi les rares travaux réalisés, Wang *et al.* [WLC03], et Mesaros et Virtanen [MV09] s'inspirent très fortement des méthodes de transcription de la parole non chantée. Le module de reconnaissance est basé sur des Modèles de Markov Cachés. Chaque phonème est modélisé par un HMM à trois états. Les probabilités d'émission de chaque état sont modélisées par des GMM, les paramètres utilisés sont des MFCC et leurs dérivées.

La différence entre les deux méthodes repose sur la phase d'apprentissage : Wang *et al.* apprennent la structure du HMM à partir de 3 200 phrases chantées, annotées en 19 600 phonèmes. Les modèles acoustiques des états du HMM sont appris sur une base de données de parole de 33 heures, en Mandarin et Taïwanais. Mesaros et Virtanen utilisent un HMM appris sur une base de données de parole (CMU ARCTIC[36]), qu'ils adaptent avec 49 extraits de 20 à 30 secondes, issus de 12 chansons pop, annotés au niveau phonétique.

Performances Wang *et al.* [WLC03] ont développé leur outil dans un but de « Query by Humming ». Il est donc naturel de le tester sur des phrases isolées : il n'y a pas d'accompagnement. Ils obtiennent ainsi un taux de reconnaissance de mots de 93 %, et de syllabes de 95 %.

Mesaros et Virtanen [MV09] évaluent leur méthode sur du chant accompagné. Les performances sont mesurées selon plusieurs critères : tout d'abord, le taux de reconnaissance de phonème, et l'accuracy (également sur les phonèmes). En testant les modèles sur les données d'apprentissage (les 49 extraits ayant servi à l'adaptation des modèles), le taux de reconnaissance est de l'ordre de 60 %, pour une accuracy de 45 %. En testant sur des données différentes, le taux de reconnaissance est de l'ordre de 40 %, pour une accuracy de 20 % [MV09]. Il est clair que de nouvelles méthodes devront être développées ; étant

[36] http://festvox.org/cmu_arctic

donné le nombre de problèmes soulevés par l'adaptation d'un transcripteur de parole pour le chant et la jeunesse des recherches, les résultats semblent prometteurs.

2.6 Conclusion

Nous avons proposé ici une brève revue des outils utilisés pour décrire divers aspects de la musique : les instruments présents, la tonalité, le tempo, le genre, les émotions, l'identité du chanteur, et les transcriptions de la mélodie principale, de la partition, de la suite d'accords et des paroles. Dans tous les travaux menés sur ces thèmes, des paramètres sont souvent créés spécifiquement pour décrire un aspect particulier de la musique. Les méthodes de classification sont par contre plus classiques.

Certaines tâches sont bien avancées ; elles sont déjà anciennes (comme la recherche du tempo) et la théorie musicale qui les accompagne est suffisamment bien construite pour pouvoir être utilisée (tonalité, accords, tempo). Des sujets restent difficiles (reconnaissance du chanteur, des instruments), tandis que d'autres, bien qu'émergeant (transcription des paroles), semblent déjà très prometteurs.

L'évaluation des méthodes pose, de nombreux problèmes. Tout d'abord, les corpora et métriques utilisés dans les divers articles, sont souvent différents. La campagne d'évaluation MIREX a permis, sur de nombreux sujets, de remédier à ce problème en mettant à la disposition de tous un cadre unifié afin d'analyser les points forts et points faibles de chaque méthode.

Se pose de plus, dans presque tous les cas, la question de la création des bases de données. Alors que certaines annotations sont à la portée de tous (genres, émotions), elles sont en revanche relativement subjectives, et demandent des annotations croisées (et indirectement d'autant plus d'annotateurs). À l'inverse, d'autres annotations sont tout à fait objectives (partition, tonalité, tempo, accords), mais elles demandent des connaissances en musique (accords, tempo, tonalité), voire même une véritable expertise (partition).

Les métriques ne sont pas toujours simples à définir. Pour la tonalité, les accords ou le tempo, une possibilité est de pondérer de différentes façons les erreurs, toute la question étant évidemment de s'accorder sur l'importance relative de chaque erreur.

Nous avons remarqué que nombre de ces tâches sont plus faciles en contexte monophonique qu'en contexte polyphonique (transcriptions, détection du tempo, reconnaissance d'instrument ou du chanteur). De plus, la détection du chant est une étape de pré-traitement nécessaire pour la reconnaissance du chanteur et la transcription des paroles.

Chapitre 3

Monophonique / Polyphonique

Sommaire

3.1 Positionnement de l'étude

Dans les tâches de description de la musique, que ce soit en vue de l'indexer ou de la transcrire, il est intéressant de connaître le nombre de sources présentes. Une source étant définie, pour nous, comme soit un chanteur, soit un instrument produisant une seule note, le nombre de sources est équivalent au nombre de notes chantées ou jouées simultanément. Cette définition est différente de celle communément employée dans la communauté « Séparation de sources » pour laquelle une source correspond à un instrument.

Des algorithmes estimations du nombre de sources harmoniques sont présents dans les algorithmes d'estimation de fréquences fondamentales multiples, les « algorithmes multipitch » [Yeh08, KNS07, Kla06, CSY+08]. Ces derniers sont basés sur des approches séquentielles : les sources sont estimées les unes après les autres. Ceci peut se faire de manière additive : on cherche à créer un signal aussi proche du signal réel, en ajoutant progressivement des sources, ou de manière soustractive : on soustrait progressivement des sources du signal, pour aboutir à un résidu aussi faible que possible. L'estimation du nombre de sources est actuellement faite *a posteriori* : les algorithmes s'arrêtent sur un critère prédéterminé ; le nombre de sources est alors égal au nombre de fréquences trouvées. Ce critère peut être :

- Un seuil sur l'énergie maximale autorisé pour le résidu (le signal auquel on a enlevé toutes les sources estimées).
- Un seuil sur l'énergie minimale de chaque source.
- Un critère sur l'information apportée par la nouvelle source estimée.

Une synthèse plus complète peut être trouvée dans la thèse de Yeh [Yeh08]).

Cependant, une connaissance *a priori* du nombre de sources présentes améliorerait les performances de ces algorithmes. En effet, on peut formuler le problème de la façon suivante : il s'agit de déterminer l'ensemble $(p_i, f_i)_{i=1..N}$ qui minimise la distance entre le signal réel et le signal reconstruit à partir des fréquences estimées, avec (p_i, f_i) le couple (puissance, fréquence) de la $i^{ème}$ sinusoïde et N le nombre de sinusoïdes. Si N est connu, le problème est évidemment beaucoup plus simple que s'il est inconnu.

Notre travail a pour but de proposer une version primaire de l'estimation du nombre de sources : il s'agit de déterminer s'il y a une ou plusieurs sources harmoniques.

3.1.1 Quelques définitions

Définissons précisément les termes « monophonie » et « polyphonie ». Dans le contexte d'analyse de la musique, un **son monophonique**, que nous appelons également **monophonie**, est défini comme un son produit par une seule source harmonique. C'est soit une note jouée par un instrument de musique, soit une note chantée par un chanteur *a capella*.

Les **sons polyphoniques**, que nous appelons aussi **polyphonies**, regroupent tous les autres sons musicaux, c'est-à-dire tous les sons produits par plusieurs sources harmoniques simultanées. Cette catégorie regroupe à la fois les orchestres, les groupes vocaux

a capella (avec plusieurs chanteurs), les chanteurs accompagnés, et les instruments polyphoniques jouant plusieurs notes simultanément. De nombreux instruments pourront être considérés soit comme monophoniques soit comme polyphoniques, selon qu'ils jouent une ou plusieurs notes simultanément (comme le piano, le violon, la guitare...) ; on peut également noter que d'autres instruments sont intrinsèquement polyphoniques, tels l'harmonica, l'accordéon, ...

Ainsi, nous considérons deux classes : « monophonie » et « polyphonie », qui peuvent être séparées en cinq sous classes : instrument solo et chanteur solo pour la classe monophonie, plusieurs instruments, plusieurs chanteurs et chanteur(s) accompagnés pour la classe polyphonie. Ces sous classes seront utilisées par la suite dans les expériences.

3.1.2 État de l'art

Jusqu'à présent, peu de travaux se sont intéressés au problème de l'estimation *a priori* du nombre de sources.

Comme résultat annexe de leur travaux sur la reconnaissance d'instruments dans le jazz, Essid *et al.* [ERB05] déterminent le nombre d'instruments (1 à 4) présents. Connaissant le genre, il est possible d'inférer les divers instruments pouvant être rencontrés. Selon les instruments reconnus, il est ensuite évident d'en donner le nombre. Le taux de bonne détection des instruments est de 91 % pour des extraits de 2 secondes. Cependant, cette méthode ne permet pas de connaître le nombre de notes jouées, puisque de nombreux instruments sont susceptibles de jouer plusieurs notes simultanément. Elle permet uniquement de donner une borne minimum. Elle est également très dépendante du style, qui est une connaissance *a priori*.

Le problème de la séparation entre « son monophonique » et « son polyphonique » a été abordé dans quelques travaux récents, à chaque fois de manière restreinte, et comme pré-traitement pour une autre tâche. Pour Smit et Ellis [SE07], il s'agit de distinguer la voix chantée solo des polyphonies vocales, pour un système de « Query by Singing » (recherche d'un morceau de musique en chantant la mélodie). Cela leur permet ensuite de ne transcrire que les parties chantées par le client du service. Pour Tsai *et al.* [TLL08], la distinction est faite entre les chanteurs solos des duos, en vue d'étendre la tâche d'identification du chanteur au cas où plusieurs chanteurs sont présents (deux dans un premier temps).

Tsai *et al.* utilisent une méthode très classique : ils modélisent la répartition des MFCC par des GMM. Ils obtiennent une accuracy de l'ordre de 96 %. Ces résultats, excellents, sont à nuancer : le problème est simplifié, puisque les polyphonies sont réduites à deux sources.

Pour leur problème, qui est plus large, Smit et Ellis prennent cette même méthode (MFCC modélisés par des GMM) comme « système de base ». Leurs résultats sont sans appel : pour une précision de 70 %, le rappel n'est que de l'ordre de 50 %.

Leur système final se base sur l'idée suivante : s'il est possible de modéliser une trame de signal avec une seule fonction périodique, il n'y a qu'un chanteur. Dans le cas contraire, il y a plusieurs instruments. Les auteurs cherchent tout d'abord la période la plus présente dans le signal, par le maximum d'autocorrélation. Après filtrage de cette fréquence, la classification est faite sur le résidu.

Un accent particulier est mis sur la construction du filtre optimal. En effet, la période à filtrer peut correspondre, dans l'idéal, à un nombre entier d'échantillons. Dans ce cas, le résidu est calculé simplement avec la formule suivante :

$$\epsilon[n] = s[n] - s[n - \tau] \tag{3.1}$$

avec τ la période, $s[n]$ le signal à l'instant n et ϵ le résidu.

Cependant, la période à filtrer peut également correspondre à un nombre non entier d'échantillons. Dans ce cas, une estimation $\hat{\mathbf{a}}$ du filtre optimal est obtenue en minimisant l'équation suivante (que nous présentons sous forme réduite) :

$$\hat{\mathbf{a}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{s} \tag{3.2}$$

avec $\mathbf{Z}_{i,j} = s[i - (\tau + j)]$ et $\mathbf{s}_i = s[i]$.

Le filtrage est ensuite réalisé ainsi :

$$\boldsymbol{\epsilon} = \mathbf{s} - \mathbf{Z}\hat{\mathbf{a}} \tag{3.3}$$

ou encore, en forme développée, avec les mêmes notations que précédemment :

$$\epsilon[n] = s[n] - \sum_{i=-k}^{k} a_i.s[n - (\tau + i)] \tag{3.4}$$

Après avoir filtré cette période, la classification est réalisée par un classifieur Bayésien sur le résidu du signal. Plus précisément, la classification est faite sur deux paramètres :

- Le ratio entre l'énergie du résidu et l'énergie du signal d'origine : cette valeur étant comprise entre 0 et 1, les auteurs utilisent des distributions Bêta pour modéliser la répartition de ce paramètre pour chacune des classes considérées.
- L'énergie (normalisée) du résidu : les auteurs modélisent les répartitions par des lois normales.

Ces deux paramètres sont traités indépendamment, la vraisemblance globale par rapport à une classe est donc le produit de la vraisemblance de chaque paramètre pour cette classe (par rapport à la Gaussienne ou à la distribution Bêta). Les modèles sont appris sur 10 minutes de signal, dont 28 % sont des solos.

Une étape de lissage est enfin ajoutée. Celle-ci est réalisée à l'aide d'un HMM à trois états. Les probabilités de transition entre les états sont apprises empiriquement en comptant le nombre de transitions dans le corpus d'apprentissage. Cela permet d'éviter des alternances rapides et non pertinentes entre classes.

Les tests sont réalisés sur 10 minutes de signal. Avec cet algorithme, pour une précision de 70 %, le rappel est de l'ordre de 70 %. Ceci représente, à précision égale, une amélioration du rappel de l'ordre de 20 % par rapport au système de base. Les auteurs notent également que cette méthode est plus fiable, puisque l'écart type des performances est divisé par deux.

Les auteurs notent que, disposant de peu de données, le système de base risque un surapprentissage, puisqu'il doit estimer 104 paramètres par classe. À l'inverse, leur méthode ne doit estimer que 4 paramètres par classe. L'annotation des données étant coûteuse, le fait d'avoir un petit corpus pour l'apprentissage est donc moins un problème.

3.2 Notre approche

Notre méthode [LAOP09b] vise à répondre à la même question de base que Tsai *et al.* : y a-t-il une ou plusieurs fréquences fondamentales ? Pour cela, nous étudions le comportement de la moyenne et de la variance d'un indice de confiance proposé par de Cheveigné et Kahawara [dCK02]. Comme nous allons le voir (partie 3.3), la distribution bivariée de ces deux paramètres est discriminante pour séparer les sons monophoniques des sons polyphoniques.

Nous traitons ce problème avec une approche probabiliste, dont l'originalité est à la fois dans le vecteur d'observation considéré, et dans sa modélisation probabiliste par des distributions de Weibull bivariées. Le système (figure 3.1) est classiquement composé de deux principaux modules : l'extraction des paramètres et la prise de décision.

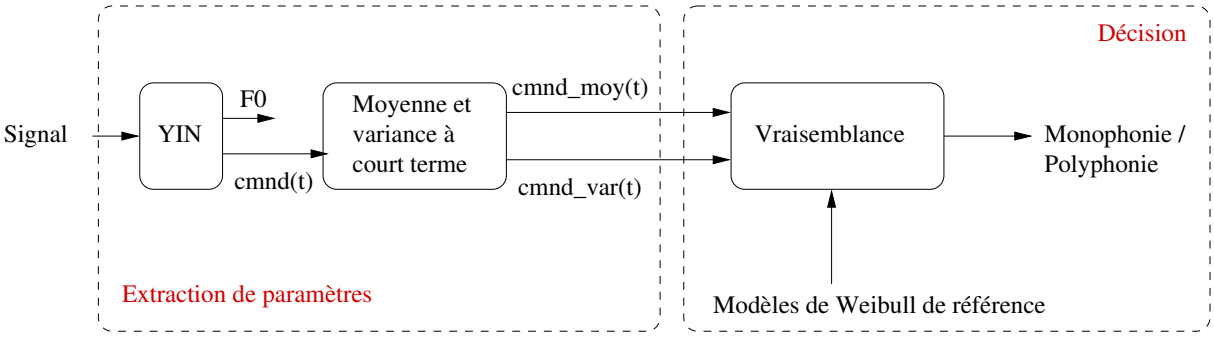


FIG. 3.1 – Schéma général de la méhode de discrimination entre sons monophoniques et polyphoniques.

3.2.1 L'extraction des paramètres

Sur chaque trame t de 10 ms (valeur classique pour l'estimation de la fréquence fondamentale), un « indice de confiance » noté $cmnd(t)$ est calculé et donne la certitude sur la valeur estimée de la fréquence fondamentale courante. Sa moyenne et sa variance à

court terme, respectivement notées $cmnd_{moy}(t)$ et $cmnd_{var}(t)$, sont calculées toutes les 10 ms, sur une fenêtre glissante centrée sur la trame t de 50 ms, soit 5 trames (valeur fixée expérimentalement), ce qui nous donne un vecteur d'observation à deux dimensions $(cmnd_{moy}(t), cmnd_{var}(t))$.

3.2.2 La prise de décision

Pour la prise de décision, nous adoptons une approche probabiliste. Nous étudions deux cas. Notre problème est un problème à deux classes. Dans un premier temps, nous considérons naturellement deux modèles, estimés sur chacune des classes (monophonie et polyphonie) : c'est l'approche « Classe ».

Les deux classes peuvent être séparées en cinq sous-classes. Ainsi, dans un second temps, les deux classes sont décrites par cinq modèles, représentant les cinq sous-classes possibles (instrument solo, chanteur solo, plusieurs instruments, plusieurs chanteurs *a capella*, instrument(s) et chanteur(s)) : c'est l'approche « Sous-Classe ». Dans tous les cas, les modèles sont des distributions de Weibull bivariées.

La décision est prise en calculant la vraisemblance des observations sur une seconde, soit 100 couples $(cmnd_{moy}(t), cmnd_{var}(t))$, par rapport à chacun des modèles. Elle est prise chaque seconde, sans recouvrement, et sans post-traitement.

3.3 L'indice de confiance - Définition et comportement statistique

3.3.1 Le YIN

Dans leur article [dCK02], de Cheveigné et Kahawara présentent une méthode d'estimation de la fréquence fondamentale d'un signal. Cet algorithme se base sur la recherche du minimum d'une fonction, la « cumulative mean normalized différence » (ou « moyenne normalisée de la différence cumulée »).

Cette fonction, dérive du calcul de la fonction de différence cumulée :

$$d_t(\tau) = \sum_{k=1}^{N} (s_k - x_{k+\tau})^2 \tag{3.5}$$

avec s le signal, N la taille de la fenêtre d'analyse, t l'indice de la trame analysée, et τ le décalage temporel.

Dans le cas d'un signal s parfaitement périodique de période T, $d_t(nT) = 0$, pour $n \in \mathbb{N}$. La période devrait donc être donnée par l'indice (non nul) du premier zéro de $d_t(\tau)$. Cependant, ceci n'est pas toujours possible, notamment si le signal audio n'est pas parfaitement périodique [dCK02]. Pour contourner ce problème, les auteurs proposent

d'utiliser plutôt la « moyenne normalisée de la différence cumulée », qu'ils définissent ainsi :

$$d'_t(\tau) = \begin{cases} 1 & si\ \tau = 0 \\ d_t(\tau) / \left[(1/\tau) \sum_{k=1}^{\tau} d_t(k) \right] & sinon \end{cases} \tag{3.6}$$

Cette nouvelle fonction modifie principalement les premières valeurs de $d_t(\tau)$ (pour τ petit). Ceci permet d'éviter que le minimum corresponde à une période trop petite, causée par du bruit haute fréquence. La période T est alors donnée par l'indice T du premier minimum de $d'_t(\tau)$.

La valeur de $d'_t(T)$ donne une idée de la fiabilité de la valeur estimée de la fréquence fondamentale : plus $d'_t(T)$ est petit, plus le signal est périodique et plus la valeur estimée de la fréquence fondamentale est fiable. Notre méthode se base sur l'étude de cet « indice de confiance », donné pour chaque trame d'analyse t. Nous l'appellerons dans la suite $cmnd(t)$.

3.3.2 Le vecteur de paramètres

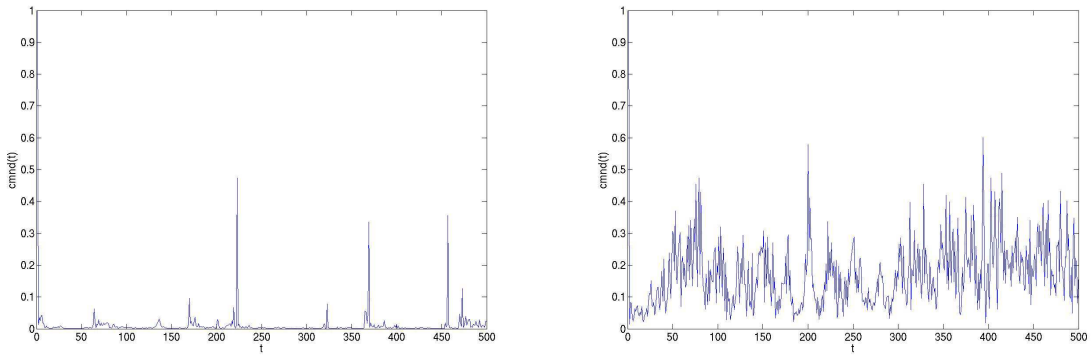
La figure 3.2 montre la différence de comportement de l'indice $cmnd(t)$ entre la musique monophonique et la musique polyphonique, sur deux extraits de 5 secondes chacun :

- Dans le cas d'un extrait monophonique, les valeurs de $cmnd(t)$ sont faibles et varient peu.
- Dans le cas d'un extrait polyphonique, elles sont élevées et varient beaucoup.

Ceci est dû au fait que, lorsqu'un seul instrument est présent, la fréquence fondamentale est bien définie, le signal est périodique. Dans ce cas, $cmnd(t)$ est toujours faible. *A contrario*, dans le cas où plusieurs instruments jouent simultanément, la fréquence fondamentale n'est pas bien définie, plusieurs périodicités sont imbriquées : $cmnd(t)$ est plus élevé. De plus, selon les instants, un instrument peut prédominer, puis aucun, puis un autre..., ce qui fait varier $cmnd(t)$.

Nous observons, pour la musique monophonique, quelques pics dans les valeurs de $cmnd(t)$. Ceux-ci correspondent à des changements de notes. En effet, l'estimateur YIN se comporte de manière semblable pour deux notes successives inclues dans une même fenêtre d'analyse et pour deux notes simultanées : aucune des deux fréquences ne peut donner un minimum faible pour la fonction de différence, l'algorithme ne trouve pas de fréquence fondamentale.

Ces remarques nous conduisent à analyser le couple (moyenne court terme, variance court terme) de $cmnd(t)$, respectivement notées $cmnd_{moy}(t)$ et $cmnd_{var}(t)$. Ces valeurs sont calculées toutes les 10 ms sur une fenêtre glissante de 50 ms, correspondant à 5 valeurs de $cmnd(t)$ (ce seuil a été déterminé expérimentalement). Les figures 3.3, 3.4, 3.5, et 3.6 montrent la répartition bivariée du couple $(cmnd_{moy}(t), cmnd_{var}(t))$ pour chaque classe et sous-classe. Les histogrammes sont calculés sur le corpus d'apprentissage, décrit



(a) Musique monophonique (b) Musique polyphonique

FIG. 3.2 – Valeurs de $cmnd(t)$ pour 5 secondes de signal.

en détail dans la partie 3.5.1. Ceci correspond à 2500 échantillons pour chaque sous-classe, 5000 pour la classe « monophonie », et 7500 pour la classe « polyphonie ».

Dans les deux cas, il y a clairement une différence de forme entre les histogrammes correspondant à la classe monophonie et ceux correspondant à la classe polyphonie. Les histogrammes des classe et sous-classes monophoniques ont leur maximum en $(0, 0)$, et sont ensuite décroissants avec une forme exponentielle. Les histogrammes des classe et sous-classes polyphoniques sont nuls en $(0, 0)$, et ont plutôt une forme Gamma : fortement croissants, avec un maximum proche du point origine, puis lentement décroissants.

3.3.3 Choix de la loi de Weibull bivariée

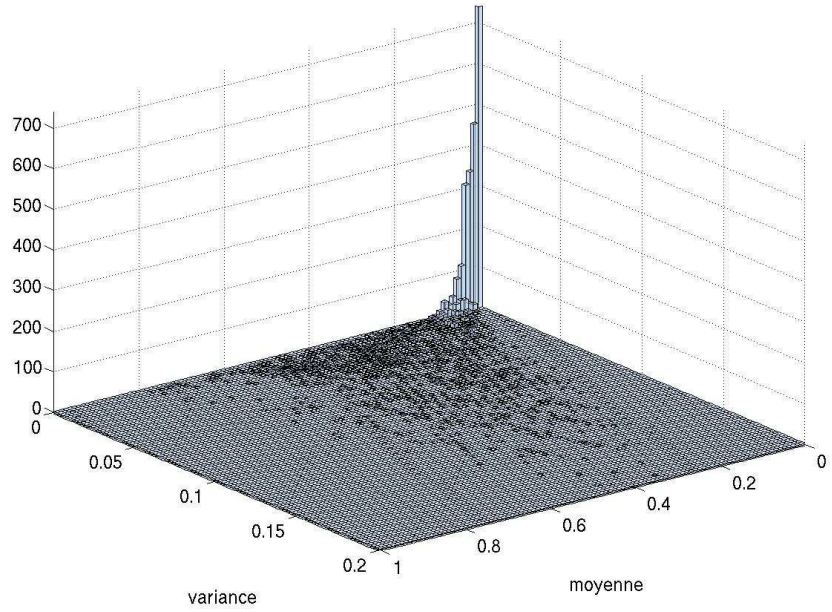
Au vu des histogrammes de répartition bivariée de nos paramètres (voir les figures 3.3, 3.4, 3.5, et 3.6), nous avons choisi de les modéliser avec des lois de Weibull bivariées. Une loi de Weibull unidimensionnelle a pour densité de probabilité :

$$f(x) = \frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1} e^{-(x/\theta)^{\beta}} \tag{3.7}$$

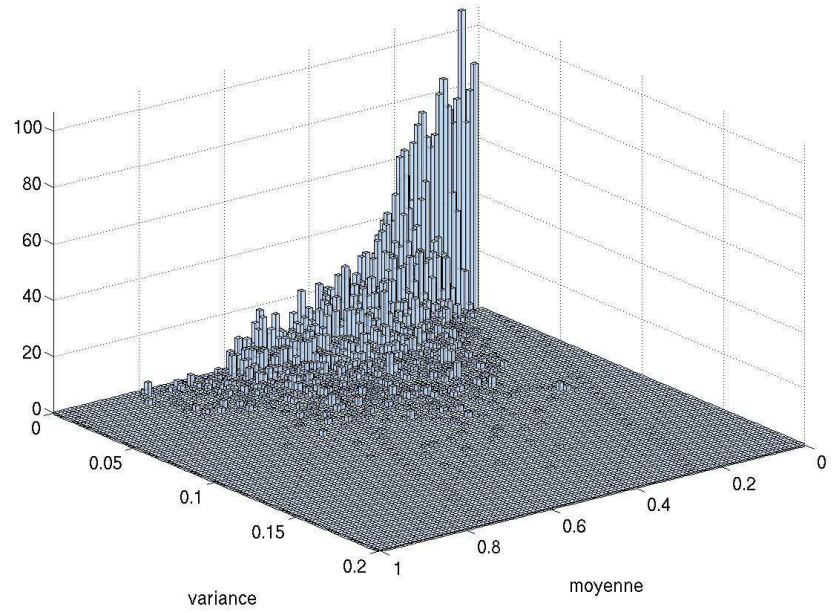
Cette loi dépend de deux paramètres : β décrit la forme de la loi et θ son échelle.

En faisant varier le paramètre de forme, il est possible d'approcher des lois très diverses (Gamma, exponentielle, Gaussienne...). De plus, elles sont efficaces pour approcher des distributions fortement dissymétriques. La figure 3.7 donne un aperçu des possibilités de cette loi.

Nous présentons dans la partie 3.3.3.1 la loi bivariée que nous allons utiliser, puis, dans la partie 3.3.3.2, nous justifions théoriquement la modélisation des histogrammes expérimentaux par des lois de Weibull bivariées.

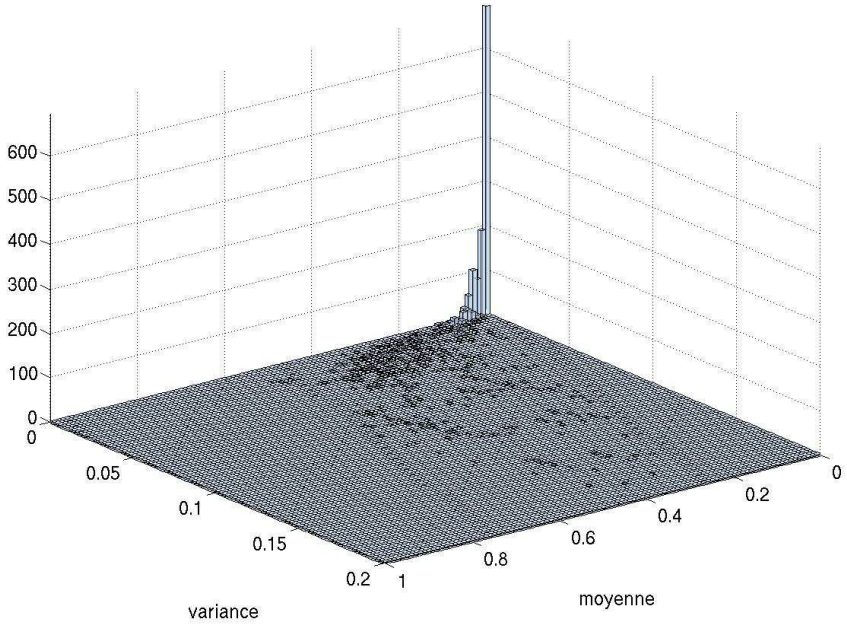


(a) Monophonies

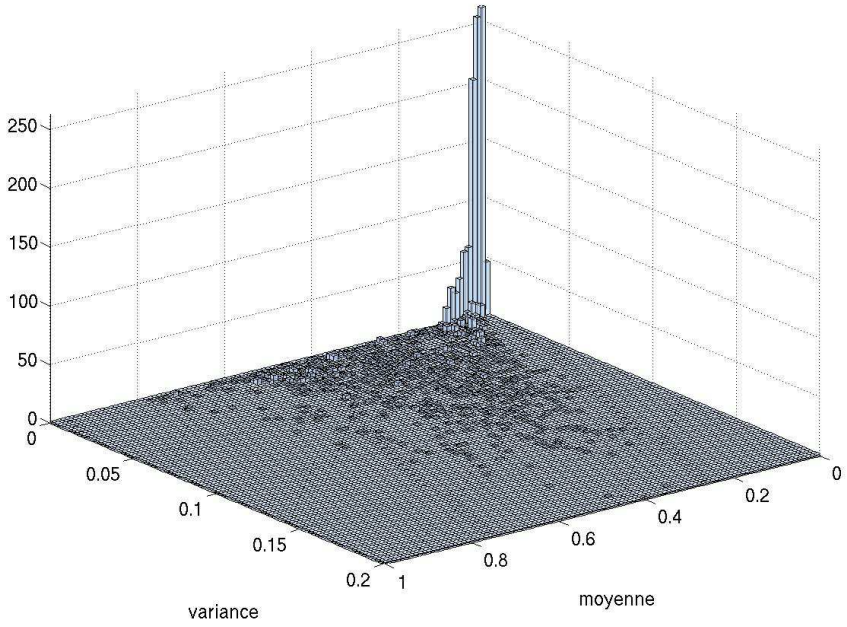


(b) Polyphonies

FIG. 3.3 – Répartition bivariée du couple $(cmnd_{moy}(t), cmnd_{var}(t))$ pour les classes monophonie et polyphonie.

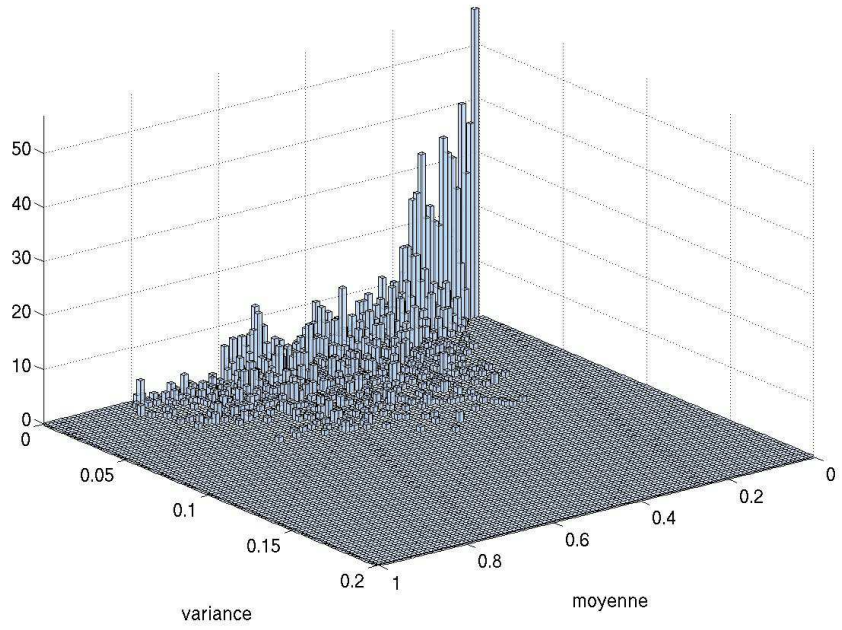


(a) Instrument solo

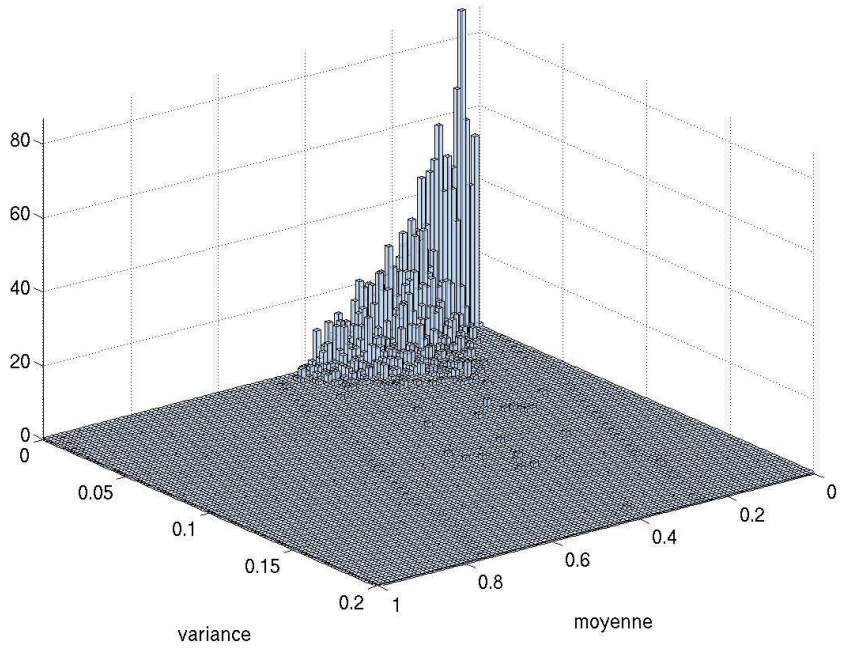


(b) Chanteur solo

FIG. 3.4 – Répartition bivariée du couple $(cmnd_{moy}(t), cmnd_{var}(t))$ pour les deux sous-classes monophoniques.



(a) Plusieurs instruments



(b) Plusieurs chanteurs

FIG. 3.5 – Répartition bivariée du couple $(cmnd_{moy}(t), cmnd_{var}(t))$ pour les deux sous-classes « Plusieurs instruments » et « Plusieurs chanteurs ».

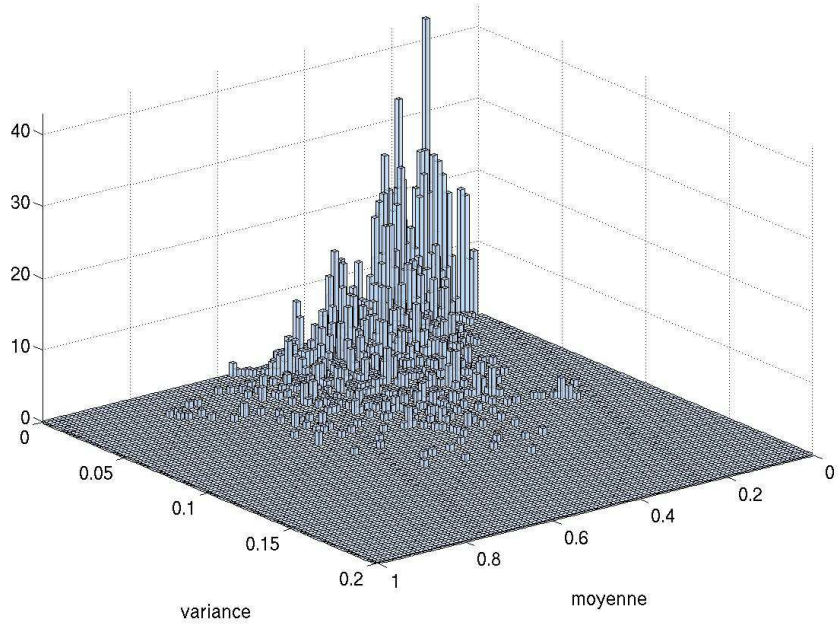


FIG. 3.6 – Répartition bivariée du couple $(cmnd_{moy}(t), cmnd_{var}(t))$ pour la deux sous-classe « Instrument(s) et chanteur(s) ».

3.3.3.1 Présentation de la loi de Weibull bivariée

La distribution de Weibull bivariée que nous utilisons a été proposée par Lu et Bhattacharyya [LB90] ; c'est une extension de la distribution proposée par Hougaard [Hou86]. La fonction de répartition est donnée par :

$$F(x,y) = 1 - \exp\left(-\left[\left(\frac{x}{\theta_1}\right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2}\right)^{\frac{\beta_2}{\delta}}\right]^{\delta}\right) \tag{3.8}$$

pour $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$, avec :

- $(\theta_1, \theta_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ les paramètres d'échelle,
- $(\beta_1, \beta_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ les paramètres de forme,
- $\delta \in]0,1]$ le paramètre de corrélation.

Pour décrire une distribution de Weibull bivariée, nous avons besoin de cinq paramètres. Une méthode d'estimation des paramètres par la méthode des moments est proposée dans la partie suivante 3.4.

3.3.3.2 Validation théorique

Pour valider théoriquement le choix d'une loi de Weibull bivariée pour modéliser les différentes répartitions, nous avons fait appel au test de Kolmogorov. Ce test, décrit en

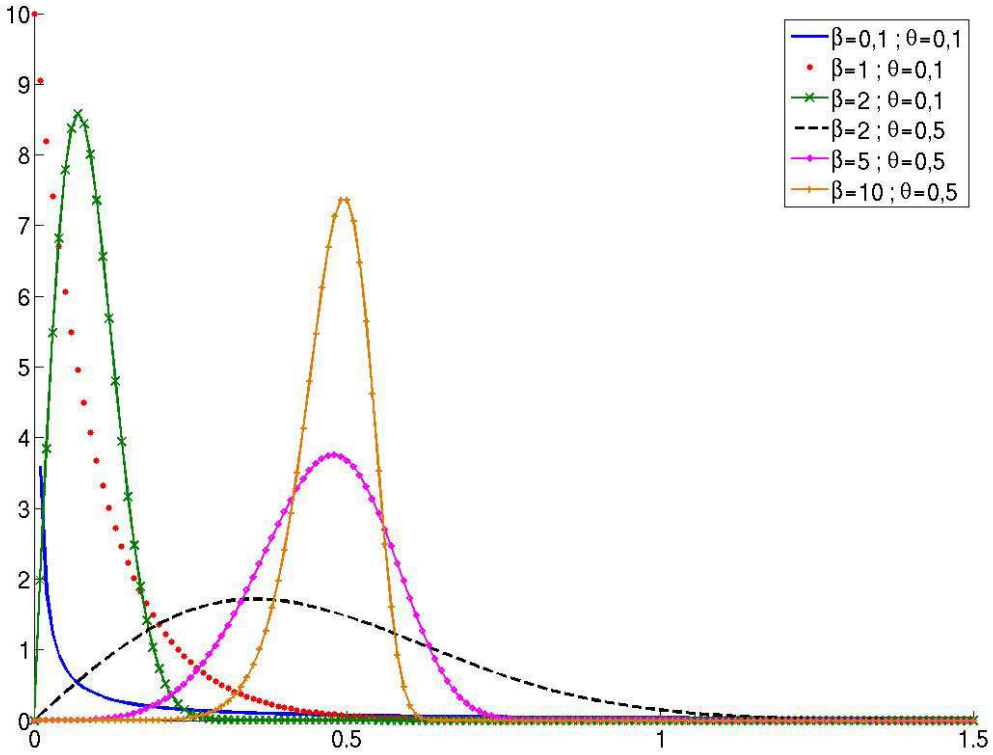


FIG. 3.7 – Densités de probabilité d'une fonction de Weibull univariée, pour différentes valeurs de paramètres d'échelle θ et de forme β.

détail dans l'annexe B a pour but de tester l'adéquation d'une distribution expérimentale à une loi de probabilité. Pour cela, il convient de mesurer l'écart maximum entre l'histogramme cumulé expérimental et la fonction de répartition théorique. Si cet écart est inférieur à un seuil, la distribution expérimentale peut être modélisée par la loi de probabilité testée.

Le seuil dépend du nombre d'échantillons disponibles pour l'estimation de l'histogramme et du risque de première espèce (risque de non détection), que nous fixons ici à 5 % (valeur classique). Les calculs théoriques ayant déjà été faits, on se réfère traditionnellement à des tables pour cette valeur. Cependant, à ce jour, les seules valeurs pré-calculées pour le seuil le sont pour des distributions unidimensionnelles. De ce fait, nous utilisons le test de Kolmogorov sur les distributions marginales.

Traditionnellement, une centaine d'échantillons est utilisée pour le test de Kolmogorov. Nous avons réalisé le test avec 105 (resp. 166) échantillons, pris aléatoirement dans le corpus d'apprentissage (décrit dans la partie 3.5.1 pour la classe monophonie (resp. la classe polyphonie). Nous avons effectué le test de Kolmogorov pour les distributions des deux paramètres (moyenne et variance) pour chacune des deux classes, pour trois lois :

la loi de Weibull, la loi normale et la loi Gamma. Le tableau 3.1 présente, pour chaque distribution marginale de chaque classe, la valeur de l'écart maximum, et le seuil auquel il faut comparer cette valeur. Le symbole $\surd$ indique que le test valide le choix de la loi, le symbole $\times$ que le test le rejette.

TAB. 3.1 – Test de Kolmogorov : valeur de l'écart maximum entre l'histogramme cumulé expérimental et la fonction de répartition théorique. Cet écart est comparé au seuil théorique. – En gras, la meilleure valeur, avec $\surd$ signifiant « accepté » et $\times$ signifiant « rejeté ».

		Weibull	Gaussienne	Gamma	Seuil théorique
Polyphonies	$cmnd_{moy}$	**0.0642** $\surd$	0.0746 $\surd$	0.0757 $\surd$	0.104
	$cmnd_{var}$	**0.0863** $\surd$	0.235 $\times$	0.107 $\times$	0.104
Monophonies	$cmnd_{moy}$	0.092 $\surd$	0.228 $\times$	**0.082** $\surd$	0.131
	$cmnd_{var}$	0.335 $\times$	**0.274** $\times$	0.358 $\times$	0.131

Pour la modélisation de la classe polyphonie, la loi de Weibull est clairement la plus adéquate. En effet, seule celle-ci a des résultats positifs au test de Kolmogorov pour les deux paramètres.

Pour la modélisation de la distribution moyenne-monophonie, deux lois sont possibles : Weibull et Gamma. La loi Gamma est plus adéquate, mais l'écart entre la loi de Weibull et la loi Gamma est faible.

Pour la modélisation de la distribution variance-monophonie, aucune loi ne satisfait le test de Kolmogorov. Les valeurs nous indiquent que la loi la moins inadéquate est la loi Gaussienne, suivie de près par la loi de Weibull.

Afin d'éviter de normaliser les scores de vraisemblance (nécessaire dès lors que l'on emploie différents types de lois), nous avons jugé pertinent de modéliser chacune des distributions marginales par des lois de Weibull, conduisant naturellement à choisir des lois de Weibull bivariées pour modéliser les distributions bivariées.

3.4 Estimation des paramètres d'une loi de Weibull bivariée par la méthode des moments

Dans cette partie, nous présentons une méthode rapide d'estimation du paramètre de corrélation de la loi de Weibull bivariée que nous utilisons. Cette méthode est dérivée de la méthode des moments.

Nous donnons tout d'abord les moments de la loi, puis nous décrivons l'estimation des paramètres par la méthode des moments [LAOP09a].

3.4.1 Les moments de la loi

Dans leur article [LB90], Lu et Bhattacharyya donnent les moments d'ordre 1 et 2 et le moment croisé d'ordre 1 de la loi bivariée que nous étudions ici :

$$E[X] = \theta_1 \Gamma\left(\frac{1}{\beta_1} + 1\right) \tag{3.9}$$

$$E[Y] = \theta_2 \Gamma\left(\frac{1}{\beta_2} + 1\right) \tag{3.10}$$

$$Var(X) = \theta_1^2 \left(\Gamma\left(\frac{2}{\beta_1} + 1\right) - \Gamma^2\left(\frac{1}{\beta_1} + 1\right)\right) \tag{3.11}$$

$$Var(Y) = \theta_2^2 \left(\Gamma\left(\frac{2}{\beta_2} + 1\right) - \Gamma^2\left(\frac{1}{\beta_2} + 1\right)\right) \tag{3.12}$$

$$Cov(X,Y) = \theta_1\theta_2 \frac{\Gamma\left(\frac{\delta}{\beta_1}+1\right)\Gamma\left(\frac{\delta}{\beta_2}+1\right)\Gamma\left(\frac{1}{\beta_1}+\frac{1}{\beta_2}+1\right) - \Gamma\left(\frac{1}{\beta_1}+1\right)\Gamma\left(\frac{1}{\beta_2}+1\right)\Gamma\left(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2}+1\right)}{\Gamma\left(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2}+1\right)} \tag{3.13}$$

3.4.2 L'estimation de θ_1, θ_2, β_1 et β_2

La détermination des valeurs de θ_1, θ_2, β_1 et β_2 est un problème d'estimation des paramètres des lois marginales de X et Y.

Les lois de Weibull univariées étant étudiées depuis de très nombreuses années (plus de 40 ans !), ce problème a déjà été très documenté, notamment par Morice [Mor68], qui résume dans son article de nombreuses méthodes pour l'estimation des paramètres.

En utilisant la méthode des moments, nous avons :

$$\frac{\Gamma(\frac{1}{\beta_1}+1)}{\sqrt{\Gamma(\frac{2}{\beta_1}+1) - \Gamma^2(\frac{1}{\beta_1}+1)}} = \frac{E[X]}{\sqrt{Var(X)}}. \tag{3.14}$$

La résolution de cette équation en β_1 se fait en utilisant des tables de valeurs pré-calculées pour les valeurs de $\Gamma(\frac{1}{\beta_1}+1)$ et de $\sqrt{\Gamma(\frac{2}{\beta_1}+1) - \Gamma^2(\frac{1}{\beta_1}+1)}$.

Nous en déduisons :

$$\theta_1 = \frac{E[X]}{\Gamma(\frac{1}{\beta_1}+1)} \tag{3.15}$$

De la même façon, nous obtenons β_2 et θ_2. Ces paramètres étant estimés, nous les considérerons connus pour l'estimation de δ.

3.4.3 L'estimation de δ

L'estimation de δ se fait indirectement par la méthode des moments, à partir de l'analyse de l'équation (3.13). Nous allons montrer que cette équation peut s'écrire sous la forme $f(\delta) = C$ avec C une constante dépendant de $Cov(X, Y)$, θ_1, θ_2, β_1 et β_2. Ainsi, δ est un zéro de $f(\delta) - C$. Puis, nous prouvons que ce zéro est unique en montrant que la dérivée $f'(\delta)$ est strictement négative pour tout $\delta \in \times]0, 1]$.

Expression de $f(\delta)$

Nous allons tout d'abord montrer que

$$(3.13) \Leftrightarrow f(\delta) = \delta B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right) = C \tag{3.16}$$

avec C une constante et $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ la fonction Bêta.

L'équation (3.13) peut également s'écrire de la façon suivante :

$$\frac{Cov(X, Y)}{\theta_1\theta_2} = \frac{\Gamma\left(\frac{\delta}{\beta_1} + 1\right)\Gamma\left(\frac{\delta}{\beta_2} + 1\right)\Gamma\left(\frac{1}{\beta_1} + \frac{1}{\beta_2} + 1\right)}{\Gamma\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2} + 1\right)} - \Gamma\left(\frac{1}{\beta_1} + 1\right)\Gamma\left(\frac{1}{\beta_2} + 1\right) \tag{3.17}$$

En posant $\Gamma\left(\frac{1}{\beta_1} + 1\right)\Gamma\left(\frac{1}{\beta_2} + 1\right) = C_1$, qui ne dépend pas de δ, et en utilisant cette propriété bien connue des fonctions Gamma : $\Gamma(a + 1) = a\Gamma(a)$, nous avons :

$$\frac{Cov(X, Y)}{\theta_1\theta_2} + C_1 = \frac{\frac{\delta}{\beta_1}\frac{\delta}{\beta_2}\Gamma\left(\frac{\delta}{\beta_1}\right)\Gamma\left(\frac{\delta}{\beta_2}\right)\Gamma\left(\frac{1}{\beta_1} + \frac{1}{\beta_2} + 1\right)}{\frac{\delta\beta_1+\delta\beta_2}{\beta_1\beta_2}\Gamma\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)} \tag{3.18}$$

En simplifiant le terme de droite par $\frac{\delta}{\beta_1\beta_2}$, nous avons :

$$\frac{Cov(X, Y)}{\theta_1\theta_2} + C_1 = \frac{\Gamma\left(\frac{1}{\beta_1} + \frac{1}{\beta_2} + 1\right)}{\beta_1 + \beta_2}\ \delta\ \frac{\Gamma\left(\frac{\delta}{\beta_1}\right)\Gamma\left(\frac{\delta}{\beta_2}\right)}{\Gamma\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)} \tag{3.19}$$

Nous remarquons que $\frac{\Gamma\left(\frac{\delta}{\beta_1}\right)\Gamma\left(\frac{\delta}{\beta_2}\right)}{\Gamma\left(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2}\right)} = B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right)$ et que $C_2 = \frac{\Gamma\left(\frac{1}{\beta_1}+\frac{1}{\beta_2}+1\right)}{\beta_1+\beta_2}$ est indépendant de δ. Ainsi, l'équation (3.13) est équivalente à l'équation suivante (avec C constant) :

$$\boxed{\delta B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right) = \frac{\frac{Cov(X,Y)}{\theta_1\theta_2} + C_1}{C_2} = C} \tag{3.20}$$

δ est donc solution de l'équation $f(\delta) - C = 0$, avec $f(\delta) = \delta B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right)$.

Signe de la dérivée de $f(\delta)$

La dérivée $f'(\delta)$ est :

$$f'(\delta) = B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right)\left[1 + \frac{\delta}{\beta_1}\left(\psi_0\left(\frac{\delta}{\beta_1}\right) - \psi_0\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)\right) + \frac{\delta}{\beta_2}\left(\psi_0\left(\frac{\delta}{\beta_2}\right) - \psi_0\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)\right)\right] \tag{3.21}$$

avec $\psi_0(x) = \frac{\mathrm{d}\ln\Gamma(x)}{\mathrm{d}x}$ la fonction digamma.

Nous savons que $B\left(\frac{\delta}{\beta_1}, \frac{\delta}{\beta_2}\right) = \frac{\Gamma(\frac{\delta}{\beta_1})\Gamma(\frac{\delta}{\beta_2})}{\Gamma(\frac{\delta}{\beta_1}+\frac{\delta}{\beta_2})} > 0$ car $\Gamma(x) > 0,\ \forall x \in \mathbb{R}^+$.

Ainsi, pour prouver que $f'(\delta)$ est strictement négative, nous devons montrer que :

$$\frac{\delta}{\beta_1}\left(\psi_0\left(\frac{\delta}{\beta_1}\right) - \psi_0\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)\right) + \frac{\delta}{\beta_2}\left(\psi_0\left(\frac{\delta}{\beta_2}\right) - \psi_0\left(\frac{\delta}{\beta_1} + \frac{\delta}{\beta_2}\right)\right) < -1 \tag{3.22}$$
$$\forall(\beta 1, \beta 2, \delta) \in \mathbb{R}^+ \times \mathbb{R}^+ \times]0, 1]$$

Nous remarquons que l'équation (3.22) ne dépend que de deux paramètres : $a = \frac{\delta}{\beta_1}$ et $b = \frac{\delta}{\beta_2}$. Ainsi, il suffit de montrer :

$$a(\psi_0(a) - \psi_0(a+b)) + b(\psi_0(b) - \psi_0(a+b)) < -1,\ \forall(a,b) \in \mathbb{R}^{+*} \times \mathbb{R}^{+*} \tag{3.23}$$

Pour cela, nous utilisons les fonctions polygamma. La fonction polygamma d'ordre n est définie comme la dérivée d'ordre n de la fonction digamma, soit la dérivée d'ordre $n+1$ du logarithme de la fonction Gamma :

$$\psi_n(x) = \frac{\mathrm{d}^{n+1}\ln\Gamma(x)}{\mathrm{d}x^{n+1}} = \frac{\mathrm{d}^n\psi_0(x)}{\mathrm{d}x^n} \tag{3.24}$$

Une propriété connue des fonctions polygamma est la relation de récurrence suivante :

$$\psi_n(x+1) = \psi_n(x) + (-1)^n n!\ x^{-(n+1)} \tag{3.25}$$

Pour $n = 1$, cela donne :

$$\psi_1(x) = \psi_1(x+1) + \frac{1}{x^2} \tag{3.26}$$

Comme $\Gamma(x)$ est convexe, nous avons $\psi_1(x) \geq 0,\ \forall x > 0$. Nous avons donc également $\psi_1(x+1) \geq 0,\ \forall x > 0$. D'où :

$$\psi_1(x) \geq \frac{1}{x^2},\ \forall x > 0 \tag{3.27}$$

Comme $\frac{1}{x^2} > 0,\ \forall x > 0$, nous avons $\psi_1(x) > 0,\ \forall x > 0$, $\psi_1(x+1) > 0,\ \forall x > 0$ et :

$$\psi_1(x) > \frac{1}{x^2},\ \forall x > 0 \tag{3.28}$$

En intégrant l'équation (3.28) entre a et $a+b$ (avec $b > 0$), il vient :

$$\psi_1(x) > \frac{1}{x^2} \Rightarrow \int_a^{a+b} \psi_1(x)\mathrm{d}x > \int_a^{a+b} \frac{1}{x^2}\mathrm{d}x \tag{3.29}$$

$$\psi_0(a+b) - \psi_0(a) > -\frac{1}{a+b} + \frac{1}{a} \tag{3.30}$$

$$a(\psi_0(a+b) - \psi_0(a)) > \frac{b}{a+b} \tag{3.31}$$

Ainsi, nous avons $a(\psi_0(a) - \psi_0(a+b)) < -\frac{b}{a+b}$.

Symétriquement, nous avons $b(\psi_0(b) - \psi_0(a+b)) < -\frac{a}{a+b}$, ce qui nous mène à l'inégalité que nous cherchions :

$$\boxed{a(\psi_0(a) - \psi_0(a+b)) + b(\psi_0(b) - \psi_0(a+b)) < -1} \tag{3.32}$$

3.5 Cadre expérimental

La tâche à laquelle nous nous intéressons ici est la distinction, au sein de la musique, entre deux classes : les sons monophoniques et les sons polyphoniques. Dans ce cadre, nous testons deux systèmes alternatifs. Dans la première expérience, nous testons la méthode telle que nous l'avons présentée : chaque classe (monophonie et polyphonie) est modélisée par une loi de Weibull bivariée. Puis nous validons notre méthode en comparant chaque étape à une méthode classique. Enfin, nous améliorons notre méthode en prenant en compte le fait que notre corpus peut être divisé en cinq sous-classes : instrument ou chanteur solo (monophonies), plusieurs instruments, plusieurs chanteurs, instrument(s) ET chanteur(s) (polyphonies).

Nous présentons tout d'abord le corpus de façon détaillée, puis la méthode d'apprentissage, et finissons par les expériences et les résultats obtenus.

3.5.1 Le corpus

Pour les expériences, nous avons utilisé un corpus « maison », contenant des observations issues de toutes les classes et sous-classes considérées dans cette étude. La liste des morceaux utilisés, ainsi que leur contenu en terme de classe / sous-classe sont donnés dans l'annexe C. Notre corpus est équilibré : les deux classes contiennent approximativement

la même quantité de données, de même que les cinq sous-classes, et ce tant pour l'apprentissage que pour les tests. Le corpus est décrit en terme de durée pour chaque classe et sous-classe dans le tableau 3.2. Dans ce tableau, nous précisons le nombre d'échantillons statistiques dont nous disposons pour l'apprentissage des lois. La décision étant prise à l'échelle de la seconde, nous précisons également le nombre de secondes de tests auxquelles correspondent les durées de test.

TAB. 3.2 – Répartition du corpus (apprentissage et test).

Classe / Sous-classe	Apprentissage		Test	
	Durée	Nb. d'éch.	Durée	Nb. de sec.
Instrument solo	25 s	2500	2 min 57 s	177
Chanteur solo	25 s	2500	5 min 38 s	338
Monophonies	**50 s**	**5000**	**8 min 35 s**	**515**
Plusieurs instruments	25 s	2500	3 min 23 s	203
Plusieurs chanteurs	25 s	2500	3 min 33 s	213
Instr. ET chanteurs(s)	25 s	2500	3 min 10 s	190
Polyphonies	**1 min 15 s**	**7500**	**10 min 6 s**	**606**
Total	**2 min 5 s**	**12500**	**18 min 41 s**	**1121**

Cinq secondes de cinq extraits différents ont été sélectionnés aléatoirement pour composer le corpus d'apprentissage. Les morceaux utilisés en apprentissage ont été totalement retirés du corpus de test. Ainsi, les genres de musiques utilisés dans la phase de test sont *a priori* différents de ceux de l'apprentissage. Le corpus étant très diversifié, nous avons en test des styles, des instruments, et des chanteurs différents de ceux utilisés lors de la phase d'apprentissage, ce qui nous permet de tester la robustesse de notre méthode. Une description détaillée du corpus pour chaque sous-classe, en terme de durée et de nombre d'extraits différents, est présentée dans les tableaux ci-dessous.

Les données que nous utilisons sont aussi variées que possible. Dans chaque classe, nous avons des styles très divers : rock, pop, musique classique, renaissance, jazz, country...

Les effectifs des orchestres, dans la sous-classe « plusieurs instruments » vont du duo à l'orchestre symphonique ou au big band en passant par des trios, des orchestres de chambre (musique classique) ou encore des petites formations (rock, jazz). Cette sous-classe est décrite plus précisément dans le tableau 3.3. Dans la sous-classe « plusieurs chanteurs » (voir tableau 3.4), nous avons également des duos, trios, petits et grands ensembles vocaux, dans des styles divers : jazz, musique moderne, opéra, country, pop... Enfin, la sous-classe « instruments et chanteurs » contient du rock, de l'opéra, du jazz, avec un ou plusieurs instruments, et un ou plusieurs chanteurs (voir tableau 3.5).

La sous-classe « instrument solo » contient des instruments tels que : le violon, le violoncelle, la contrebasse, la flûte à bec, la guitare, la clarinette, le triangle, le piano et

TAB. 3.3 – Description de la sous-classe « plusieurs instruments ».

Style	Apprentissage		Test	
	Durée	Nb. d'extraits	Durée	Nb. d'extraits
Jazz	5 s	1	37 s	3
Pop	5 s	1	35 s	4
Musique moderne	5 s	1	24 s	2
Renaissance	5 s	1	15 s	1
Baroque	5 s	1		
Classique			44 s	5
Rock			35 s	3
Country			13 s	1
Total	**25 s**	**5**	**203 s**	**19**

TAB. 3.4 – Description de la sous-classe « plusieurs chanteurs ».

Sexe	Apprentissage		Test	
	Durée	Nb. d'extraits	Durée	Nb. d'extraits
Hommes et Femmes	25 s	5	105 s	6
Hommes			68 s	3
Femmes			40 s	2
Total	**25 s**	**5**	**213 s**	**19**

TAB. 3.5 – Description de la sous-classe « instrument(s) et chanteur(s) ».

Style	Apprentissage		Test	
	Durée	Nb. d'extraits	Durée	Nb. d'extraits
Jazz	5 s	1	16 s	1
Pop	10 s	2	42 s	6
Musique moderne	5 s	1	21 s	2
Opéra	5 s	1	30 s	2
Rock&Roll			10 s	2
Rock			22 s	2
Country			33 s	3
RAP			16 s	2
Total	**25 s**	**5**	**190 s**	**20**

TAB. 3.6 – Description de la sous-classe « instrument solo ».

Instruments	Apprentissage		Test	
	Durée	Nb. d'extraits	Durée	Nb. d'extraits
Flûte à bec	10 s	2	31 s	3
Contrebasse[37]	10 s	2	19 s	4
Violon[37]	5 s	1	22 s	2
Triangle			3 s	1
Guitare[37]			17 s	3
Hautbois			21 s	3
Clarinette			18 s	2
Piano[37]			17 s	3
Trompette			10 s	2
Violoncelle[37]			19 s	3
Total	**25 s**	**5**	**177 s**	**26**

la trompette (voir tableau 3.6). Les instruments potentiellement polyphoniques (piano, cordes) jouent bien une seule note à la fois. Enfin, la sous-classe « chanteur solo » contient des extraits de douze chanteurs différents, hommes et femmes, professionnels et amateurs (voir tableau 3.7).

TAB. 3.7 – Description de la sous-classe « chanteur solo ».

Sexe	Apprentissage			Test		
	Nombre de chanteurs	Durée	Nombre d'extraits	Nombre de chanteurs	Durée	Nombre d'extraits
Hommes	1	5 s	1	5	3 min 20 s	5
Femmes	3	20 s	4	3	2 min 18 s	3
Total	**4**	**25 s**	**5**	**8**	**5 min 38 s**	**8**

Notons que tous les styles, instruments, et effectifs ne sont pas présents dans le corpus d'apprentissage.

3.5.2 L'apprentissage

Dans un premier temps, nous considérons les deux classes de notre problème : les sons monophoniques et les sons polyphoniques. Dans un deuxième temps, ces deux classes sont subdivisées en cinq sous-classes : instrument solo, chanteur solo, plusieurs instruments,

[37] Jouant une seule note

plusieurs chanteurs, instruments et chanteurs. Ainsi, sept modèles sont appris, un pour chaque classe et chaque sous-classe.

Pour chaque sous-classe, le modèle est appris avec 25 secondes de signal (5 secondes de 5 extraits musicaux différents). $cmnd_{moy}$ et $cmnd_{var}$ étant calculés toutes les 10 ms, 25 secondes de signal correspondent à 2500 couples pour l'estimation des cinq paramètres de la distribution de Weibull bivariée.

Pour chaque classe, les modèles sont appris avec toutes les données des sous-classes correspondantes : 50 secondes de 10 extraits différents pour la classe monophonie, 75 secondes de 15 extraits différents pour la classe polyphonie, correspondant respectivement à 5000 et 7500 échantillons pour l'estimation des lois de Weibull bivariées. Ainsi, si les durées d'apprentissage semblent faibles, elles n'en correspondent pas moins à un nombre raisonnable d'échantillons pour estimer une loi de probabilité.

Les figures 3.8 et 3.9 montrent les différentes distributions de Weibull bivariées estimées pour chaque classe et sous-classe. Sur la figure 3.8, nous rappelons les histogrammes bivariés expérimentaux, à titre de comparaison. Tout comme pour les histogrammes de répartition, les figures représentant les monophonies sont très différentes de celles représentant les polyphonies.

3.6 Résultats expérimentaux

Une première expérimentation est réalisée pour évaluer le système basé sur l'approche « Classe ». Les résultats sont donnés dans la partie 3.6.1.

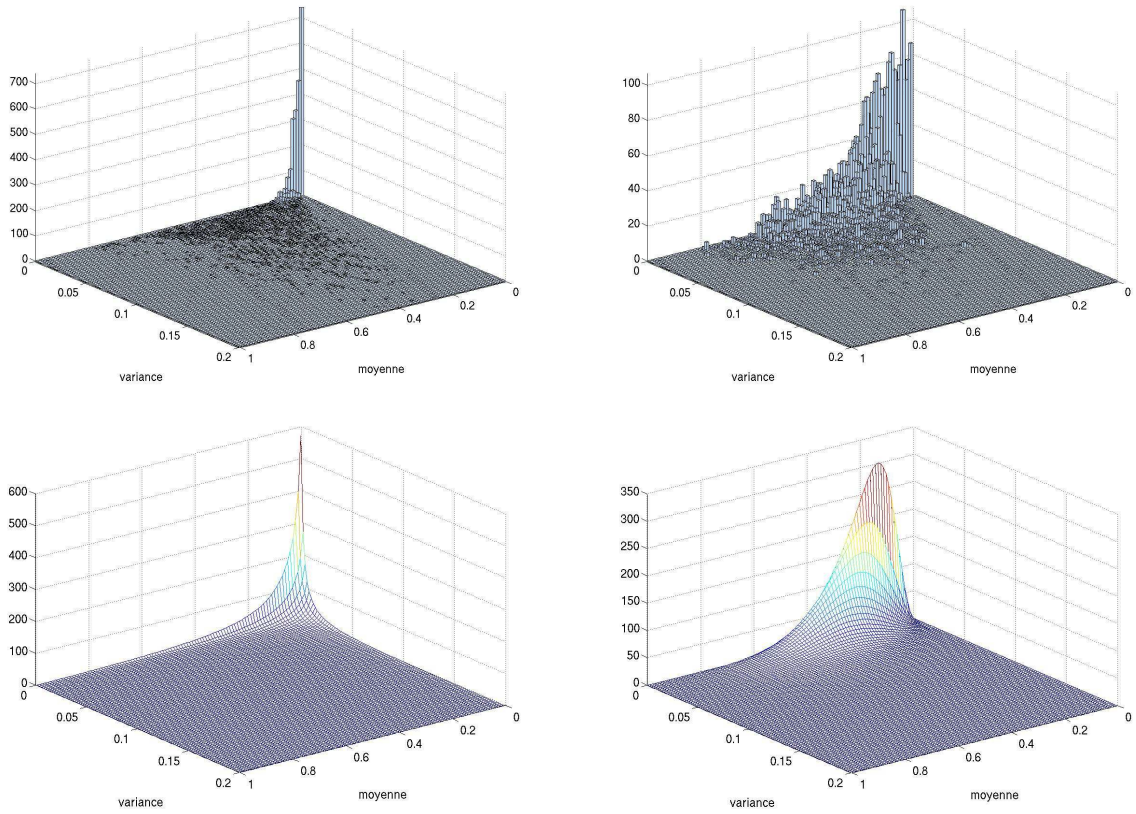
Dans la partie 3.6.2, nous validons chacune des étapes (paramétrisation, modélisation, classification), en comparant les performances à des méthodes dites « classiques » :

- Un système de base : la paramétrisation est faite avec des MFCC, la modélisation par des GMM, et la classification par le maximum de vraisemblance.
- Un système « Gaussien » : la paramétrisation est faite avec nos paramètres, la modélisation avec des lois Gaussiennes bivariées, la classification par le maximum de vraisemblance.
- Un système « Weibull univarié » : la paramétrisation est faite avec nos paramètres, la modélisation avec des lois de Weibull univariées, la classification par le maximum de vraisemblance.
- Un système « SVM » : la paramétrisation est faite avec nos paramètres, la classification avec des SVM.

Enfin, dans la partie 3.6.3, nous proposons une amélioration de notre méthode avec l'introduction des « Sous-classe ».

Dans toutes ces alternatives, nous avons utilisé le même corpus, avec la même division de celui-ci entre apprentissage et test. Les performances sont mesurées à l'aide du taux global d'erreur :

$$err = \frac{Nombre\ de\ secondes\ mal\ classifiées}{Nombre\ total\ de\ secondes}$$



(a) Monophonies (b) Polyphonies

FIG. 3.8 – En haut, les histogrammes bivariés expérimentaux. En bas, les distibutions de Weibull bivariées estimées pour chacune des deux classes.

Pour chaque expérience, nous donnons également la matrice de confusion.

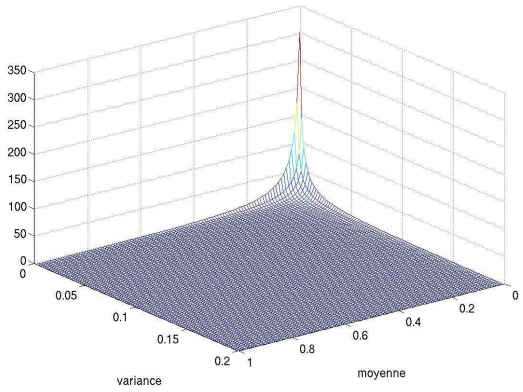
3.6.1 Le système primaire : l'approche « Classe »

La répartition des paramètres est modélisée par une loi de Weibull bivariée pour chaque classe (monophonie et polyphonie) : c'est l'approche « Classe ». Le taux d'erreur est de **8,5±1,6 %**. La matrice de confusion est donnée dans le tableau 3.8.

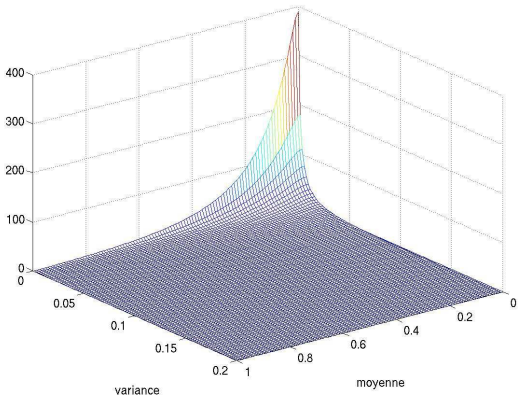
TAB. 3.8 – Matrice de confusion pour l'approche « Classe ».

	Monophonie	Polyphonie
Monophonie	**82,1 %**	17,9 %
Polyphonie	1,3 %	**98,7 %**

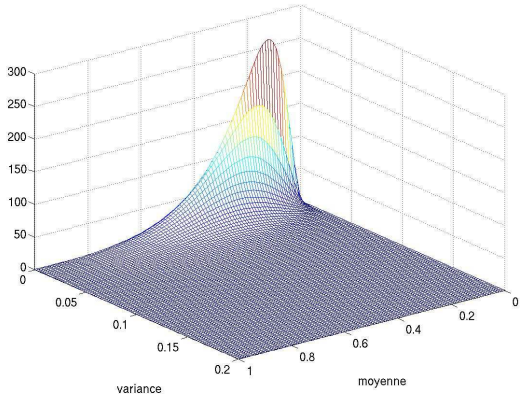
Nous remarquons qu'il y a plus d'erreurs dans la classe monophonie. Ces erreurs sont dues à des chanteurs ou instruments solos jouant sur un tempo rapide. L'algorithme YIN



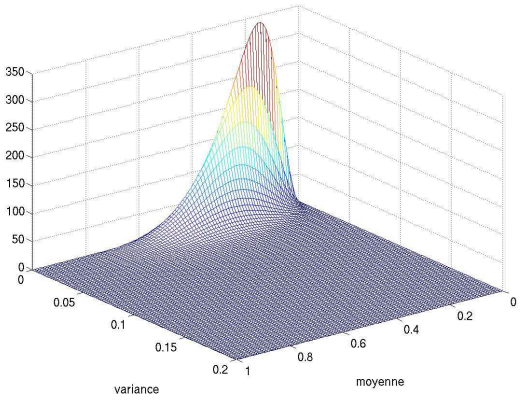
(a) Instrument solo



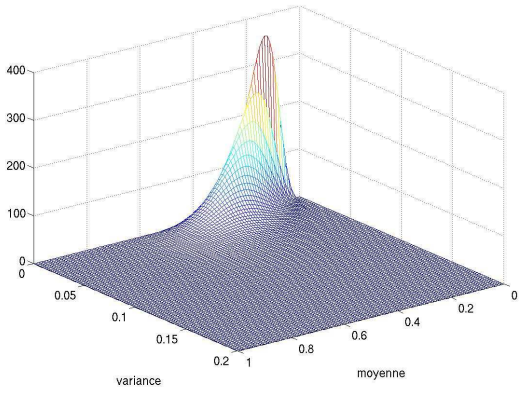
(b) Chanteur solo



(c) Plusieurs instruments



(d) Plusieurs chanteurs



(e) Instument(s) et chanteur(s)

FIG. 3.9 – Distributions de Weibull bivariées estimées pour les cinq sous-classes.

a besoin d'une durée minimale pour estimer la fréquence fondamentale. Si la note est trop courte, l'estimation est faite en considérant deux notes consécutives, ce qui biaise le

résultat : l'estimateur se comporte comme s'il y a deux notes, $cmnd(t)$ est donc élevé. Ces erreurs sont également dues au fait que les chanteurs prononcent des consonnes, qui sont non voisées. Dans ces zones, l'estimateur YIN ne peut pas trouver une fréquence fondamentale qui n'existe pas, $cmnd(t)$ se comporte là aussi comme en contexte polyphonique.

À l'inverse des instants de musique polyphonique ont été reconnus comme de la monophonie car le son est à cet instant particulièrement harmonique. Par exemple, il peut s'agir d'un accord parfait, dans lequel les notes jouées sont les harmoniques les unes des autres.

Nous remarquons que le fait qu'il y ait (voir le descriptif détaillé du corpus) des instruments, styles, et chanteurs présents dans le corpus de test qui ne l'étaient pas dans le corpus d'apprentissage n'affecte pas le résultat final.

3.6.2 Comparaison avec des méthodes classiques - Validation de la méthode proposée

3.6.2.1 Système de base

Dans un premier temps, nous comparons notre méthode au schéma généralement utilisé comme « système de base » en traitement des données audio. La paramétrisation est faite à l'aide de MFCC et la modélisation avec des GMM.

Nous avons testé différentes configurations, tant pour la paramétrisation que pour la modélisation. Pour la paramétrisation, nous avons retenu comme paramètres : soit l'Énergie et 12 coefficients MFCC (13 paramètres), soit l'Énergie et 12 coefficients MFCC, auxquels nous ajoutons ensuite leurs dérivées (26 paramètres). Dans chaque cas, nous avons fait varier le nombre de composantes pour les GMM de 1 à 256, avec des matrices de covariance diagonales. Pour le GMM à une composante, nous avons également testé la configuration dans laquelle la matrice de covariance est pleine.

Le tableau 3.9 présente les résultats obtenus pour chaque configuration. Dans notre contexte, la meilleure configuration consiste à paramétriser le signal en utilisant les dérivées, et à modéliser la répartition des paramètres avec un GMM à 16 composantes. On obtient alors un taux d'erreur de **19,2±2.3 %**. La matrice de confusion pour cette configuration est présentée dans le tableau 3.10.

Tab. 3.9 – Taux d'erreur pour les différentes configurations testées (en %).

Nb de GMM	1	2	4	8	16	32	64	128	256	1 (MP)[38]
E+12 MFCC	45,4	50,3	37,2	34,9	33,4	34,4	32,2	30,9	28.9	46,1
E 12 MFCC+Δ	22,6	25,8	22	20,8	**19,2**	21,3	20	32,6	27,1	38,9

[38]MP : Matrice pleine

TAB. 3.10 – Matrice de confusion pour le système de base.

	Monophonie	Polyphonie
Monophonie	**88,3 %**	11,7 %
Polyphonie	25,4 %	**74,6 %**

La première conclusion, la plus évidente, est que cette méthode n'est pas appropriée. Le fait que le taux d'erreur ne change que peu (avec les deux paramétrisations), pour des GMM avec 4 à 64 composantes montre que la paramétrisation n'est pas optimale pour séparer les deux classes.

Nous notons cependant que l'ajout des dérivées améliore très nettement les résultats, qui passent en moyenne de 30 % à 20 % d'erreur.

Enfin, les piètres résultats obtenus avec un grand nombre (128 ou 256) de composantes pour les GMM doivent être dus à la faible taille du corpus d'apprentissage. Nous n'avons en effet que 5000 (resp. 7500) vecteurs pour apprendre le modèle de la classe monophonie (resp. polyphonie).

Nous remarquons que ces résultats sont cohérents avec ceux de Smit et Ellis [SE07], qui obtenaient une F-mesure de 70 % (voir partie 3.1.2).

3.6.2.2 Validation des paramètres et de la modélisation

Dans cette expérience, le système testé est le suivant : les paramètres sont ceux que nous avons proposés dans cette étude ($cmnd_{moy}$ et $cmnd_{var}$), et la modélisation est faite à l'aide de lois Gaussiennes bivariées (avec des matrices de covariance pleines).

Le taux global d'erreur est de **11,4±1.8 %**, la matrice de confusion est présentée dans le tableau 3.11.

TAB. 3.11 – Matrice de confusion en utilisant deux modèles Gaussiens bivariés.

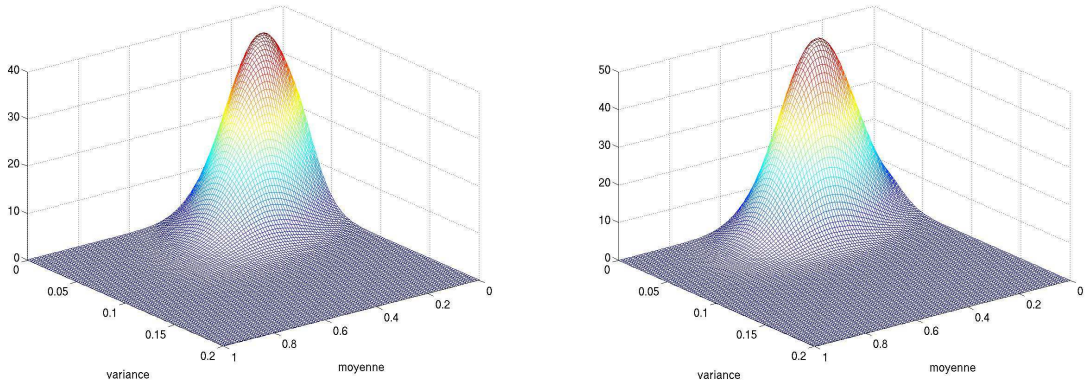
	Monophonie	Polyphonie
Monophonie	**89,1 %**	10,9 %
Polyphonie	11,7 %	**88,3 %**

Cette expérience nous permet de valider notre méthode sous plusieurs aspects.

Tout d'abord, elle valide notre paramétrisation. Cette méthode peut être vue comme le système de base (partie 3.6.2.1), dans lequel nous aurions gardé la modélisation est par des lois Gaussiennes bivariées, et changé la paramétrisation. De cette manière, le taux d'erreur est presque divisé par deux !

Ensuite, cette expérience valide expérimentalement le choix des lois de Weibull pour la modélisation. Cette méthode peut également être vue comme une modification de celle

que nous avons proposée (partie 3.6.1), dans laquelle nous aurions gardé la même paramétrisation, mais changé la modélisation par des lois Gaussiennes. Les lois normales bivariées estimées à partir du corpus d'apprentissage sont présentées sur la figure 3.10.



(a) Monophonies (b) Polyphonies

FIG. 3.10 – Distibutions normales bivariées estimées pour chacune des deux classes.

Deux explications nous semblent pertinentes pour expliquer ce résultat :

- Nous confirmons expérimentalement le résultat que nous avons montré dans la partie 3.3.3.2 : les lois de Weibull sont plus adaptées pour la modélisation des distributions expérimentales que nous avons.
- Lois normales modélisent le fait que les moyennes et variances des distributions sont différentes, mais ne sont pas capables de modéliser leur principale différence : leurs formes. Les lois de Weibull, elles, nous permettent de modéliser correctement le fait que la majorité des valeurs sont concentrées en (0,0) pour les monophonies, avec une forme de type exponentielle, et de modéliser dans le même temps une forme plus proche d'une Gamma ou d'une Gaussienne pour les polyphonies, sans valeurs au point (0,0).

Il semble donc qu'une bonne part de l'information réside dans la forme spécifique de chacune des distributions.

3.6.2.3 Validation de l'approche bivariée

Dans cette expérience, tout comme dans la précédente, les paramètres sont ceux que nous avons proposés dans cette étude ($cmnd_{moy}$ et $cmnd_{var}$). La modélisation est faite, pour chaque classe, à l'aide de deux lois de Weibull univariées indépendantes, une pour chaque paramètre.

Le taux global d'erreur est de **15,5±2.1 %**, la matrice de confusion est présentée dans le tableau 3.12.

TAB. 3.12 – Matrice de confusion. Chaque classe est modélisée par des modèles de Weibull univariés indépendants.

	Monophonie	Polyphonie
Monophonie	**85,3 %**	14,7 %
Polyphonie	16,2 %	**83,8 %**

Nous validons ici le choix d'une approche bivariée. Le fait de prendre en compte la corrélation entre les deux paramètres, et de modéliser leur répartition avec une loi bivariée divise presque par deux le taux d'erreur.

Si on se reporte à la figure 3.2, qui présente les courbes d'évolution de $cmnd(t)$, il semble bien que la moyenne et la variance sont plus corrélées dans le cas monophonique que dans le cas polyphonique.

3.6.2.4 Validation de l'approche probabiliste

Pour valider notre choix d'une approche probabiliste, nous comparons notre méthode à une autre approche bien connue pour traiter de problèmes à deux classes : les Machines à Vecteurs de Support, ou SVM. Nous avons testé trois noyaux : Gaussien, Polynomial et Sigmoïde. Pour chacun de ces noyaux, nous avons cherché les paramètres optimaux (paramètre du noyau et paramètre de régularisation). Pour pouvoir comparer une approche SVM, qui classe chaque vecteur indépendamment des autres, et la nôtre, qui prend une décision sur 100 vecteurs adjacents, nous avons adopté la stratégie suivante :

- classer chaque vecteur indépendamment (approche naïve),
- pour 100 décisions consécutives (une seconde), voter à la majorité (approche « à la seconde »).

Les performances obtenues pour chaque noyau, ainsi que le nombre de vecteurs support sélectionnés sont résumés dans le tableau

TAB. 3.13 – Comparaison des performances et nombre de vecteurs supports obtenus pour les différents noyaux SVM

Noyau	Gaussien	Sigmoïde	Polynomial
Taux global d'erreur	23 %	41,8 %	56,7 %
Nombre de vecteurs supports	8330	8329	7235

La meilleure configuration est obtenue avec un noyau Gaussien, avec pour paramètres optimaux $C = 1$ pour le paramètre de régularisation, et $\sigma = 0,7$ pour l'écart-type du noyau Gaussien. Le taux global d'erreur est de **22,5±2.4 %**, la matrice de confusion est présentée dans le tableau 3.14.

Tab. 3.14 – Matrice de confusion - SVM à 2 classes, avec un vote à majoritaire sur 1 seconde.

	Monophonie	Polyphonie
Monophonie	**50,6 %**	49,4 %
Polyphonie	3,9 %	**96,1 %**

Ces résultats, très surprenants au premier abord, le sont moins lorsqu'on examine de près les vecteurs d'apprentissage sélectionnés comme supports : dans chaque expérience, plus de 7000 (sur 12500) vecteurs (voir le tableau 3.13) ont été sélectionnés ! Plus précisément, dans la classe monophonie, tous les vecteurs (sauf 2 !) sont sélectionnés comme supports. Ainsi, la capacité de généralisation de la classe « monophonie » par le SVM est nulle, ce qui explique que la classification de cette classe soit aléatoire. En revanche, relativement peu de vecteurs sont sélectionnés comme supports (environ 2000) pour représenter la frontière de la classe polyphonie. La projection dans l'espace des attributs ne semble pas rendre les classes séparables : tout se passe comme si la classe « monophonie » est incluse dans la classe « polyphonie ».

L'approche naïve, consistant à classer chaque vecteur indépendamment des autres, ne peut donc pas marcher. Une approche « à la seconde », en utilisant un vote à la majorité sur la classification de 100 vecteurs consécutifs n'apporte que très peu d'améliorations (de l'ordre de 0.5 % sur le taux d'erreur). Ceci est probablement dû aux résultats trop aléatoires de la première étape de cette méthode.

3.6.3 Une amélioration : l'approche « Sous-classe »

Compte tenu de la composition de notre corpus, en deux classes et cinq sous-classes, nous proposons de modéliser la répartition des paramètres pour chaque sous-classe par une loi de Weibull bivariée : c'est l'approche « Sous-classe ». La fusion est ensuite faite de manière évidente : les sous-classes « instrument solo » et « chanteur solo » sont fusionnées dans la classe « monophonie », les trois autres sous-classes dans la classe « polyphonie ». Le taux d'erreur est de **6,3±1.4 %**, la matrice de confusion est présentée dans le tableau 3.15.

Tab. 3.15 – Matrice de confusion pour l'approche « Sous-classe ».

	Monophonie	Polyphonie
Monophonie	**88,7 %**	11,3 %
Polyphonie	4,8 %	**95,2 %**

Le fait de considérer cinq modèles pour l'apprentissage et la classification, avec une étape de fusion des cinq sous-classes en deux classes améliore très sensiblement les résultats. On observe un peu plus d'erreurs dans la classe polyphonie mais un taux d'erreur

beaucoup plus faible dans la classe monophonie. Ceci mène à une diminution de 2 % du taux d'erreur global.

Une étude détaillée des résultats pour chaque sous-classe et pour chaque configuration (deux et cinq modèles) est présentée dans le tableau 3.15. Nous remarquons que les deux sous-classes de la classe « monophonie » sont mieux classées, avec une très nette amélioration pour la sous-classe « chanteur solo », alors que quelques erreurs supplémentaires sont faites dans chacune des sous-classes de la classe « polyphonie ».

TAB. 3.16 – Résultats pour chaque sous-classe - 2 et 5 modèles.

	5 modèles		2 modèles	
	Monophonie	Polyphonie	Monophonie	Polyphonie
Instrument solo	**79,5 %**	20,5 %	**70,2 %**	29,8 %
Chanteur solo	**90,7 %**	9,3 %	**79,4 %**	20,6 %
Plusieurs instruments	6 %	**94 %**	0,6	**99,4 %**
Plusieurs chanteurs	7,1 %	**92,9 %**	4,3	**95,7 %**
Instrument(s) et chanteur(s)	2,1 %	**97,9 %**	0,6	**99,4 %**

3.7 Conclusion

Nous avons présenté un outil de classification entre les sons monophoniques et les sons polyphoniques. Cet outil se base sur la modélisation par des lois de Weibull bivariées de la moyenne et de la variance à court terme d'un indice de confiance.

Nos contributions se situent dans deux domaines. Au niveau *traitement du signal*, nous avons proposé deux nouveaux paramètres, qui nous semblent pertinents pour notre tâche. Nous avons également proposé d'utiliser une distribution inhabituelle dans ce domaine : la loi de Weibull.

Dans le *domaine probabiliste*, nous avons proposé une méthode d'estimation des paramètres d'une loi de Weibull bivariée par la méthode des moments. Les valeurs de ces paramètres peuvent être prises telles quelles, ou servir de bonne initialisation à un algorithme d'optimisation pour les valeurs des paramètres.

Des comparaisons avec diverses méthodes nous ont permis de valider les différentes étapes de notre méthode : la paramétrisation, la modélisation par des lois de Weibull bivariées, et l'approche probabiliste.

Notre approche donne de très bons résultats : le taux global d'erreur est de 6,3 %, contre 19,2 % pour une approche « état de l'art ». Notre méthode s'avère très rapide : pour un fichier d'une minute, le temps d'exécution est de l'ordre de 12 secondes.

Cependant, de nombreux travaux restent à faire. Il s'agira d'améliorer notre outil, en nous attaquant aux principaux problèmes, à savoir :

- dans les monophonies, les successions rapides de notes, qui sont actuellement classés comme de la polyphonie,
- dans les polyphonies, les accords « trop parfaits », qui sont actuellement classés comme de la monophonie.

Enfin, notre méthode traite actuellement chaque seconde indépendamment des autres. Il est tout à fait possible d'envisager un post-traitement, qui permettrait d'enlever les décisions aberrantes, et améliorerait *de facto* les résultats. Un post-traitement réellement adapté ne pourra en revanche être mis en œuvre que dans le cadre d'une application bien déterminée. De même, le choix de l'approche « Classe » ou de l'approche « Sous-classe » pourra éventuellement dépendre de l'application envisagée, en fonction de la répartition du corpus entre les « Sous-classe ».

Chapitre 4

Détection du chant

Sommaire

4.1 Introduction

La question que nous nous posons ici est la suivante : **« Dans un morceau de musique donné, y a-t-il du chant, et si oui, à quels instants ? »**

La détection automatique du chant par l'analyse de la bande audio est un problème relativement récent, les premières recherches sur le sujet datent d'une dizaine d'années. Deux communautés se sont principalement intéressées à ce sujet.

D'une part, la communauté « musique » recherche un outil de description de la musique très précieux. De ce point de vue, ces travaux sont dans la continuité de ceux réalisés pour la reconnaissance des instruments. Ils sont également indispensables comme pré-traitement pour d'autres tâches telles que l'identification du chanteur, ou encore la transcription des paroles de chansons.

D'autre part, la communauté « signal », après s'être intéressée à la détection de la parole et de la musique, s'est naturellement penchée sur le problème du chant. De ce point de vue, le chant peut effectivement être vu comme un son « intermédiaire » entre de la parole « pure » et de la musique « pure » (instrumentale).

Le chant a en effet bel et bien des caractéristiques qui le rapprochent de la parole de par son mode de production : il est évidemment produit tout comme elle par la voix humaine, et il peut transmettre un message. Mais il a aussi des caractéristiques qui le rapprochent de la musique de par son contexte d'utilisation, il est souvent accompagné d'instruments, il a une mélodie comme support. La plupart des systèmes de classification Parole / Musique le classent d'ailleurs dans la catégorie Musique.

Une troisième communauté s'intéresse aux caractéristiques du chant : la communauté « synthèse ». Celle-ci cherche les caractéristiques à ajouter sur la mélodie brute pour la faire ressembler à du chant. Certaines études dans ce domaine peuvent faire apparaître des paramètres caractéristiques (notamment le vibrato), dont nous nous inspirons dans notre étude.

Une étude intéressante menée par Ohishi *et al.* [OGIT05] étudie la capacité de l'être humain à discriminer le chant de la parole. La conclusion est qu'il faut perceptuellement un signal d'une durée d'au moins une seconde pour avoir un taux de reconnaissance de 100 %. Entre 0,5 et 1 seconde, le taux de reconnaissance est encore supérieur à 90 %, mais en deçà de 0,5 seconde, le taux chute dramatiquement.

Nous présentons tout d'abord un état de l'art des recherches menées sur la détection du chant (partie 4.2), puis dans la partie 4.3 nous décrivons les outils (paramètres et segmentations) que nous utilisons. Dans la partie 4.4, nous présentons la détection du chant ; cette détection est différenciée en fonction du contexte monophonique ou polyphonique du signal. Enfin, la partie 4.5 est consacrée aux expériences menées et aux résultats obtenus.

4.2 État de l'art

Les travaux effectués sur le chant diffèrent à plusieurs titres : paramétrisation, modélisation et corpus d'étude.

4.2.1 Les paramètres utilisés

S'inspirant des travaux fructueux réalisés en détection de la parole et de la musique, de nombreuses études se fondent sur les MFCC. Lukashevich *et al.* [LGD07] ont testé les modèles GMM appris sur des MFCC. Les résultats varient entre 72 % et 81 % en terme de F-mesure. Les MFCC ont par ailleurs été évalués par Rocamora et Herrera [RH07] comme étant la meilleure caractéristique pour détecter la voix chantée. La comparaison est effectuée en utilisant la plupart des paramètres classiques en reconnaissance de parole, que nous avons présentés dans la partie 2.2.1. En n'utilisant que les MFCC comme paramètres, avec une modélisation par des SVM (Support Vector Machines), les performances atteignent 76,6 % de bonne classification pour des segments de une seconde. Ces expériences montrent que si les MFCC peuvent être des paramètres intéressants pour la détection du chant, ils ne peuvent pas être utilisés seuls. La bonne stratégie est probablement celle utilisée par Markaki *et al.* [MHS08] : utiliser les MFCC pour améliorer un système existant. Ils améliorent effectivement le score de 2 % (passant de 11 % à 9 % de taux d'erreur).

Les autres paramètres habituellement utilisés en complément des MFCC sont décrits ci-dessous.

Comme dans de nombreux domaines en traitement du signal, nous avons les paramètres « classiques », que nous avons présentés dans la partie 2.2 : le centroïde spectral, les moments du spectre, le flux spectral [RH07, RRD08], l'énergie [Zha03], le taux de passage à zéro [RH07, RRD08], l'énergie par bande de fréquence [MM06], les LPCC [BE01, RH07, RRD08], les ondelettes [MM06].

D'autres paramètres ont été créés spécifiquement pour la détection du chant. Nous remarquons que tous ces paramètres visent à étudier les fréquences présentes, et plus particulièrement leur stabilité et leur évolution. En effet, la fréquence fondamentale de la voix chantée est connue pour ne pas être parfaitement stable (nous détaillons ce point dans la partie 4.3.1).

Maddage *et al.* [MWXW04] proposent la « Twice-Iterated Composite Fourier Transform », qui permet d'analyser fréquentiellement la Transformée de Fourier. Markaki *et al.* [MHS08] utilisent une représentation semblable du signal : la « Modulation Frequency Analysis » (Analyse de la modulation de fréquence). Ce paramètre analyse la modulation d'amplitude dans chaque bande de fréquence.

Chou et Gu [CG01] utilisent un nouveau paramètre, le coefficient harmonique, que nous avons présenté dans la partie 2.2.2.

De nombreux paramètres sont basés sur une analyse de la fréquence fondamentale F_0. Ohishi *et al.* [OGIT05] segmentent par exemple la fréquence fondamentale en « Note Like Unit », et modélisent sa trajectoire sur chacun des segments par des trigrams. Santosh *et al.* [SRRR09] étudient la stabilité des harmoniques du signal pour distinguer le chant des instruments.

Il ne faut pas oublier dans cet inventaire le vibrato, qui a été utilisé comme paramètre caractéristique dès les premiers travaux [Ger02], et qui l'est encore actuellement [RP09, KNL08, NL07a]. Nous présentons ce paramètre en détail dans la partie 4.3.1.

Enfin, récemment, la recherche du chant en contexte polyphonique a mené certains auteurs [SRRR09, RP09] à étudier les partiels contenus dans le signal, afin d'isoler ceux appartenant au chant – une telle approche avait été initiée par Taniguchi *et al.* [TAO+05] avec la segmentation sinusoïdale, que nous utilisons dans ce travail et présentons dans la partie 4.3.2.1.

4.2.2 Méthodes de classification

Berenzweig et Ellis [BE01] utilisent les Modèles de Markov Cachés (HMM : Hidden Markov Models), très performants pour la transcription de la parole. En partant de l'idée que la parole et le chant partagent certaines caractéristiques, notamment la structure formantique et les transitions entre phonèmes, ils utilisent les probabilités *a posteriori* apprises sur un modèle de parole comme paramètre pour la détection du chant. Les HMM sont également utilisés pour le post-traitement [RRD08]. Ils permettent en effet de prendre en compte la durée du chant.

Ce problème peut être vu comme un problème de classification à deux classes. De nombreux travaux utilisent ainsi des méthodes de décision Bayésiennes, en modélisant la répartition des paramètres par des outils également très utilisés dans la communauté « parole » : les GMM [LGD07, TZW08, ER02]. Enfin, d'autres travaux utilisent les SVM (Machines à Vecteur de Support) [RRD08, RH07] ou encore les k-Plus Proches Voisins [RH07].

4.2.3 Les corpora étudiés, les résultats obtenus

Nous remarquons que beaucoup de travaux se restreignent à des corpora particuliers. Dans les premières études [Ger02, OGIT05], les auteurs se plaçaient naturellement dans un contexte monophonique.

Puis les corpus se sont élargis, par exemple en considérant un accompagnement percussif [KR03]. Aujourd'hui, de nombreuses études portent spécifiquement sur le chant dans la musique Pop [MWXW04, NSW04, KNL08, RH07].

D'autres études utilisent des enregistrements « faits maison », par exemple un ensemble de personnes parlant et chantant la même phrase [MM06, OGIT05]. Ces travaux sont centrés sur l'étude des différences entre la parole et le chant.

Quelques études [TAO+05, ER02] se veulent plus exhaustives, en intégrant dans leur corpus du jazz, de la musique moderne, de la country, ou encore de la musique classique.

Ainsi, il est difficile de comparer les différentes méthodes – d'autant plus que les métriques utilisées sont également variées. Nous avons cependant noté que les résultats bruts (obtenus sans post-traitement) sont actuellement de l'ordre de 75 à 80 % de bonne détection, quel que soit le corpus considéré.

Pour cette étude, nous utilisons le même corpus que pour la séparation monophonie / polyphonie. Celui-ci, détaillé dans la partie 3.5.1, a été construit de façon à être aussi diversifié que possible, tout en gardant un certain équilibre entre les classes. Ainsi, nous avons des exemples relativement simples, notamment dans les deux sous-classes instrument solo et chanteur solo, mais nous explorons également les exemples polyphoniques, avec là des exemples qui peuvent être extrêmement difficiles, notamment dans la musique contemporaine, le jazz ou encore du chant accompagné d'orchestres.

4.3 Les paramètres de notre étude

La méthode que nous avons développée se base principalement sur la détection du vibrato. Nous présentons cette notion dans la partie 4.3.1. Celle-ci étant définie à partir de la fréquence fondamentale, elle n'est pas directement utilisable pour l'analyse de musique polyphonique. À l'aide de deux segmentations (la segmentation sinusoïdale et la segmentation pseudo-temporelle), que nous décrivons dans la partie 4.3.2, nous étendons cette notion au cas polyphonique, à l'aide de la notion de *vibrato étendu*, que nous présentons dans la partie 4.3.3.

4.3.1 Le vibrato

4.3.1.1 Définition

Seashore [Sea38] définit le vibrato comme « *a pulsation of pitch, usually accompanied with synchronous pulsations of loudness and timbre, of such extent and rate as to give a pleasing flexibility, tenderness, and richness to the tone* », ce que nous pouvons traduire comme « *une oscillation de la fréquence fondamentale, habituellement accompagnée de variations synchrones de la puissance et du timbre, dont l'étendue et la fréquence sont telles qu'elles ajoutent au son une flexibilité, une tendresse et une richesse plaisante* ».

Le vibrato est donc une oscillation périodique de la fréquence fondamentale d'un instrument ou d'un chanteur. La plupart des auteurs distinguent deux types de vibratos : l'oscillation se fait soit sur la valeur de la fréquence, soit sur l'intensité. Quand l'oscillation se fait sur l'intensité du son, ce vibrato est parfois appelé « trémolo » [RP09, VGD05]. Notons que ces deux types ne sont pas exclusifs l'un de l'autre – pour certains instruments

et mécanismes de production, ces deux types sont même indissociables. Dans le cas des instruments, le vibrato est un effet musical ajouté volontairement par le musicien.

4.3.1.2 Mécanismes de production

Les mécanismes de production du vibrato sont différents selon l'instrument considéré. Timmers et Desain [TD00] en décrivent certains.

Sur les instruments à cordes (violons), le vibrato est produit par l'oscillation périodique de la main gauche sur le manche de l'instrument, autour de la position correspondant à la fréquence fondamentale de la note. Dans ce cas, seule la valeur de la fréquence fondamentale change, l'intensité étant maintenue constante si la pression de l'archet sur les cordes ne change pas.

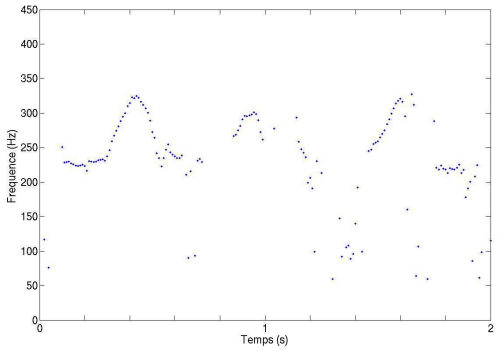
Sur la plupart des instruments à vent (flûtes, hautbois, cuivres), le vibrato peut être produit par une modulation du souffle, réalisée à l'aide du diaphragme et de la gorge : de manière générale, souffler plus fort fait légèrement monter la note. Avec ce mode de production, il y a variation de l'intensité et de la fréquence de la note. Sur les bois, il peut également être produit par la fermeture et l'ouverture partielle d'un trou de l'instrument [Vib09]. Cette méthode était à l'origine appelée le « flattement ». Dans ce cas, l'intensité du souffle restant constante, le vibrato ne change que la hauteur de la note.

Pour le chant, les mécanismes de production sont moins connus. Il semblerait [Sun94] qu'ils dépendent de la culture. Dans la musique classique occidentale (dans l'opéra principalement), il serait produit par des contractions du muscle crico-thyroïdien, qui se trouve dans le larynx et qui participe au contrôle de la tension des cordes vocales. Dans la musique Pop et la musique non occidentale, il proviendrait souvent d'une variation de la pression sous-glottale (variation de la pression du souffle).

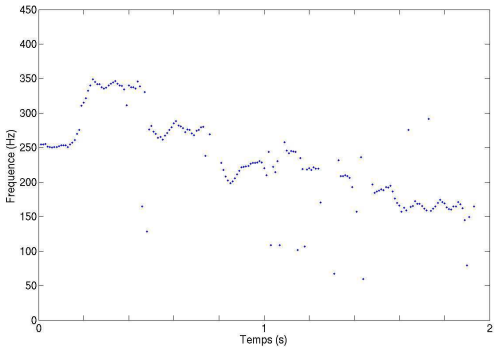
4.3.1.3 Caractéristiques du vibrato des chanteurs

Le vibrato du chant est de type « oscillation de la valeur de la fréquence fondamentale ». La première de ses caractéristiques est qu'il est produit spontanément par la voix humaine [TD00]. L'origine de ce vibrato spontané est très discutée [AC07]. Toujours est-il que les auteurs sont cependant d'accord sur le fait que, même si les chanteurs professionnels peuvent dans une certaine mesure le contrôler (l'atténuer ou au contraire l'amplifier), il est toujours présent lorsque quelqu'un chante [TD00, AC07, Sea38]. La deuxième caractéristique intéressante est la fréquence des oscillations : pour les instruments, cette fréquence est choisie par le musicien. Pour le chant, cette oscillation est toujours à un rythme compris entre 4 et 8 Hz (ces valeurs varient selon les auteurs : 6,5 Hz [Sea38], 5 Hz [AC04] 6 à 12 Hz [Ger02], 5 à 7 Hz [Sun94, MH00], 4 à 6,7 Hz [RDS$^+$99]). Si le rythme des oscillations est souvent considéré comme constant pour un chanteur donné [Sea38, Sun94, NL07b], et est même utilisé pour l'identification du chanteur [NL07b], l'étendue fréquentielle des oscillations est très variable, même pour un chanteur, et peut aller jusqu'à plus d'un demi-ton (140 *cents* [MH00]).

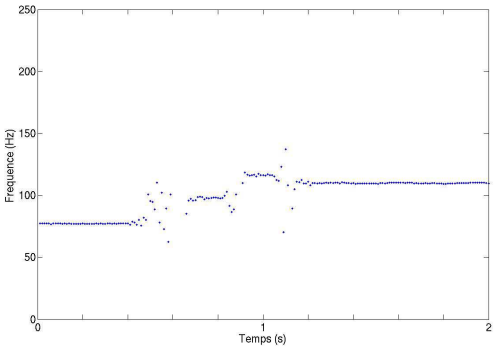
La figure 4.1 montre la fréquence fondamentale d'une personne (une femme) qui parle, d'une personne (la chanteuse Barbara) qui chante et d'un instrument de musique (contrebasse). La fréquence fondamentale de la parole est instable, celle de la musique est au contraire très constante note par note. La fréquence fondamentale du chant présente clairement du vibrato sur chaque note : elle oscille périodiquement autour d'une valeur centrale.



(a) Parole



(b) Chant



(c) Musique

FIG. 4.1 – Fréquence fondamentale d'une personne qui parle (a), d'une personne qui chante (b) et d'un instrument de musique (c). Il n'y a du vibrato que pour le chanteur.

Sur un suivi de fréquence F donné, nous choisissons de suivre la méthode de Gerhard [Ger02] : il y a du vibrato si la transformée de Fourier de F présente un maximum entre 4 et 8 Hz.

4.3.2 Une segmentation du signal

Dans notre travail, nous analysons des extraits musicaux monophoniques et polyphoniques. Dans le cas d'extraits polyphoniques, la notion de « fréquence fondamentale du

signal » n'a pas de sens, puisque les différents instruments/chanteurs présents ont chacun une fréquence fondamentale.

Cependant, les harmoniques étant des multiples de la fréquence fondamentale, s'il y a du vibrato, celui-ci est présent à la fois sur la fréquence fondamentale et sur ses harmoniques. Dans cette partie, nous présentons une méthode de suivi temporel des fréquences présentes dans le signal. Nous analysons ensuite chacun des suivis pour déterminer s'il contient du vibrato.

4.3.2.1 La segmentation sinusoïdale

Cette segmentation, développée par Taniguchi *et al.* [TAO+05] repose sur une analyse à la fois temporelle et fréquentielle du signal. Elle réalise le suivi temporel des fréquences. Un segment sinusoïdal est défini par quatre paramètres :

- l'indice de début,
- l'indice de fin,
- le vecteur des fréquences,
- le vecteur des amplitudes.

La taille de ces deux vecteurs est déterminée par la durée du segment.

Dans leur article, les auteurs proposent la méthode suivante pour la recherche des segments sinusoïdaux :

1. Calculer le spectre (par exemple toutes les 10 ms, avec une fenêtre de Hamming de 20 ms).
2. Lisser le spectre (par exemple un filtre moyenneur sur trois points).
3. Convertir les fréquences en *cent* (100 *cent* = 1/2 *ton*) de la manière suivante :

$$f_{cent} = 1200.log_2\left(\frac{f_{Hz}}{440.2^{\frac{3}{11}-5}}\right). \tag{4.1}$$

4. Détecter les maxima du spectre : pour la trame t, nous avons deux ensembles $(f_t^i)_{i=1..N}$ les fréquences et $(p_t^i)_{i=1..N}$ les amplitudes, avec N le nombre de maxima du spectre. Dans notre cas, nous avons fixé une limite à N, en nous basant sur un échantillon représentatif de signaux sonores harmoniques. Expérimentalement, nous avons constaté que pour ces signaux, N est typiquement de l'ordre de 15. Nous avons donc imposé $N < 40$ (nous prenons en compte au maximum les 40 premiers maxima du spectre), afin d'être sûrs de trouver tous les maxima.
5. Calculer les distances entre les différents maxima du spectre de deux instants successifs :

$$d_{i_1,i_2}(t) = \sqrt{\left(\frac{f_t^{i_1} - f_{t-1}^{i_2}}{C_f}\right)^2 + \left(\frac{p_t^{i_1} - p_{t-1}^{i_2}}{C_p}\right)^2} \tag{4.2}$$

   Avec C_f et C_p deux constantes.

6. Relier les points entre eux : deux points $(t, f_t^{i_1})$ et $(t+1, f_{t+1}^{i_2})$ appartiennent au même segment sinusoïdal si $d_{i_1,i_2}(t) < d_{th}$. C_f, C_p et d_{th} sont déterminés expérimentalement : $C_f = 100$ (1 demi-ton), $C_p = 3$ (puissance divisée par 2) et $d_{th} = 5$. Pour C_f et C_p, nous avons utilisé les valeurs proposées par Taniguchi *et al.* [TAO+05]). Pour la valeur de d_{th}, nous utilisons une valeur différente de celle proposée dans [TAO+05], notre étude ayant montré une meilleure adéquation de cette valeur à un corpus varié.

La figure 4.2 présente un exemple de segmentation sinusoïdale pour un extrait de 23 secondes de chant monophonique *a capella*. Les différents segments sinusoïdaux sont visiblement les harmoniques les uns des autres. La zone **A** (comprise entre 0 et 3 secondes), dans laquelle cela n'est pas le cas, correspond à une zone de bruit (une respiration). Lorsque du vibrato est présent sur la fréquence fondamentale (le segment sinusoïdal le plus bas), il se retrouve également sur tous les segments sinusoïdaux qui en sont les multiples (les harmoniques).

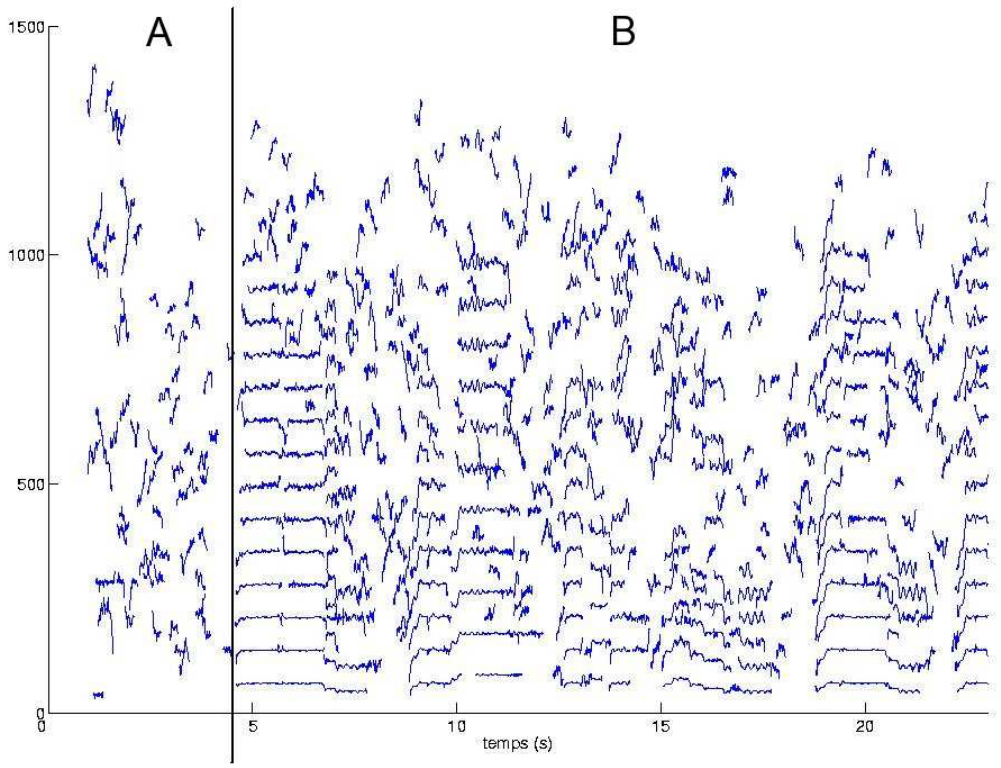


FIG. 4.2 – Segmentation sinusoïdale d'un extrait de 23 secondes de chant monophonique *a capella* : chaque ligne (bleue) est un segment.

4.3.2.2 La segmentation pseudo-temporelle

Dans notre travail, nous voulons étudier la présence de vibrato sur les harmoniques. Cela suppose, dans l'idéal, de grouper les harmoniques note par note. Si des études sont actuellement à l'œuvre sur la question, notamment pour des problématiques d'estimation de fréquences fondamentales multiples, elles ne sont pas encore assez abouties pour nos besoins.

Nous proposons [LAOP07] une alternative, qui est de grouper temporellement les harmoniques dont les limites (débuts et fins) sont temporellement corrélées.

Cette segmentation, que nous appelons « pseudo-temporelle », est réalisée à partir de la segmentation sinusoïdale. Il s'agit de :

1. Trouver toutes les extrémités temporelles des segments sinusoïdaux, en distinguant les débuts des fins.
2. Placer une limite de segment pseudo-temporel à la trame t s'il y a au moins 2 extrémités à t ET 3 débuts ou 3 fins entre t et $t+1$.

Un segment pseudo-temporel est alors défini par deux limites successives. On distingue immédiatement deux types de segments :

- Les segments longs et stables (durée supérieure à 100 ms). Dans le cas d'un son monophonique, ils correspondent à une note ; dans le cas d'un son polyphonique, ils correspondent au mieux à un accord, sinon à une zone stable harmoniquement (sans changement de note).
- Les segments courts. Ils correspondent aux zones de transition, le temps que toutes les harmoniques des notes « sortent » : transition entre deux notes pour un son monophonique, transition entre deux accords pour un son polyphonique.

La figure 4.3 présente un exemple de segmentation pseudo-temporelle pour le même extrait de chant monophonique *a capella* que sur la figure 4.2.

4.3.3 Le vibrato étendu

Avec la notion de segment pseudo-temporel, nous sommes maintenant en mesure d'élargir le concept de vibrato : c'est le « vibrato étendu » [LAOP07]. Ce nouveau paramètre mesure, dans un segment pseudo-temporel, la proportion de segments sinusoïdaux qui ont du vibrato. Ce nouveau paramètre sera noté *vibr*. Comme nous l'avons décrit dans la partie 4.3.1, les segments provenant du chant ont du vibrato, ainsi, *vibr* sera plus élevé en présence de chant.

Pour pouvoir décider de la présence de vibrato sur un suivi de fréquences (fréquence fondamentale ou segment sinusoïdal), il est nécessaire que ce suivi dure un certain temps. Ainsi, nous ne prenons pas en compte les segments sinusoïdaux trop courts (dont la durée est inférieure à 50 ms).

De la même manière, les segments pseudo-temporels courts sont des segments de transition, dans lesquels par définition il n'y a pas de note fixe. Nous ne cherchons donc pas la

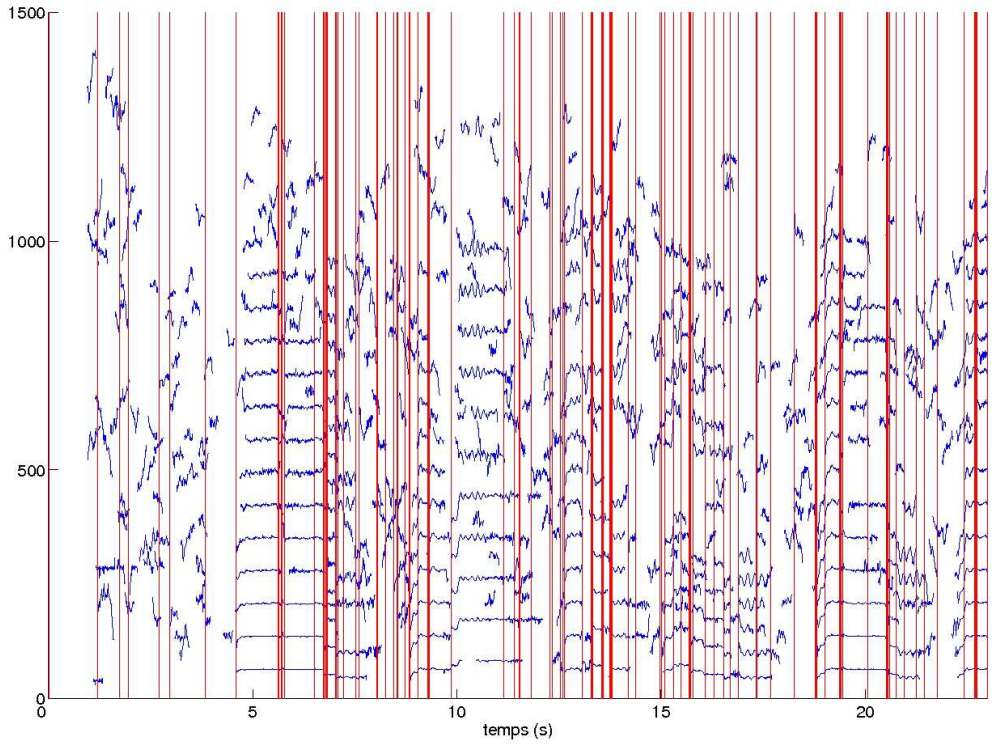


FIG. 4.3 – Segmentation temporelle du même extrait que la figure 4.2, les lignes verticales sont les limites des segments.

présence de vibrato sur les segments pseudo-temporels courts (durée inférieure à 50 ms). Pour ces segments, nous attribuons la valeur $vibr = 0$ (il n'y a pas de vibrato).

Pour les segments pseudo-temporels longs, *vibr* est calculé de la façon suivante :

$$vibr = \frac{\sum_{s \in \Gamma} l(s)}{\sum_{s \in \Omega} l(s)} \tag{4.3}$$

avec :

$l(s)$ la longueur du segment s,

Γ les segments sinusoïdaux longs (>50 ms) avec du vibrato,

Ω l'ensemble des segments sinusoïdaux longs.

4.4 La détection du chant

Dans cette partie, nous présentons les différentes méthodes que nous avons développées pour la détection du chant. Dans un premier temps, nous présentons l'approche « pri-

maire », dans laquelle la classification est réalisée en se basant sur la valeur du vibrato étendu. Puis, dans la partie 4.4.2, nous précision la définition de l'expression « présence de chant ». Cette nouvelle définition, ainsi que la prise en compte du contexte (monophonique ou polyphonique), nous permet d'adapter notre méthode de décision. Nous présentons cette amélioration dans la partie 4.4.3.

4.4.1 Le système primaire

Afin de mesurer la pertinence des paramètres que nous avons proposés (segmentation pseudo-temporelle et vibrato étendu), nous avons tout d'abord construit un système de détection du chant basé uniquement sur ceux-ci [LAOP07]. Le système est présenté sur la figure 4.4.

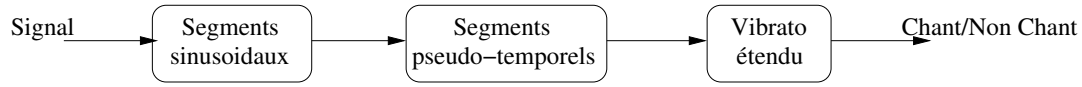


FIG. 4.4 – Schéma général du système primaire.

Nous pouvons estimer que chaque segment sinusoïdal correspond réellement à une harmonique d'un instrument présent, et que les différents segments sinusoïdaux proviennent de différents instruments. Nous espérons que quelques-uns des segments sinusoïdaux détectés correspondent aux harmoniques du (des) chanteur(s) s'il y a du chant.

La décision est prise en ne se fondant que sur le vibrato étendu : nous analysons plus précisément la moyenne du vibrato étendu sur 1 seconde.

Du chant est détecté si la valeur moyennée de *vibr* sur une seconde est supérieure à un seuil, que nous avons expérimentalement fixé à $0,08$ (voir l'expérience décrite dans la partie 4.5.2).

4.4.2 Une nouvelle définition du chant

Notre étude nous a fait prendre conscience d'une caractéristique du chant : un chanteur ne chante pas tout le temps, il fait souvent des pauses. Il y a ainsi souvent des pauses très courtes, inférieures à 0,5 secondes, notamment pour les respirations. Dans le cadre polyphonique, il y a aussi des pauses courtes, allant jusqu'à 3 secondes, avec typiquement des transitions instrumentales, et des pauses longues, qui peuvent aller jusqu'à une minute ou plus, pour des parties instrumentales.

Autant, lors des pauses longues, il est pertinent de considérer qu'il n'y a pas de chant, autant lors des pauses courtes et très courtes la question se pose. Dans notre travail nous considérons qu'il n'y a pas d'interruption du chant pour des pauses inférieures à 1 seconde.

De ces considérations, nous tirons une nouvelle définition du chant :

> Une seconde de signal est considérée comme étant chantée si du chant est perceptible pendant une durée non nulle.

Cette nouvelle définition nous conduit à changer de méthode de décision : la décision est toujours prise pour chaque seconde, mais la détection du chant se base maintenant sur le fait que, durant un certain temps, nous avons détecté la présence de vibrato ou un du vibrato étendu non nul.

4.4.3 Prise en compte du contexte monophonique ou polyphonique

Afin d'améliorer notre système primaire, nous proposons de réaliser un pré-traitement : pour chaque seconde de signal, nous déterminons tout d'abord s'il y a une ou plusieurs sources harmoniques. Ceci nous permet ensuite d'adapter notre processus de décision à chaque contexte. Cette adaptation est réalisée en tenant compte de la nouvelle définition du chant que nous avons proposée ci-dessus [LAOP09c].

Le système global est résumé sur la figure 4.5.

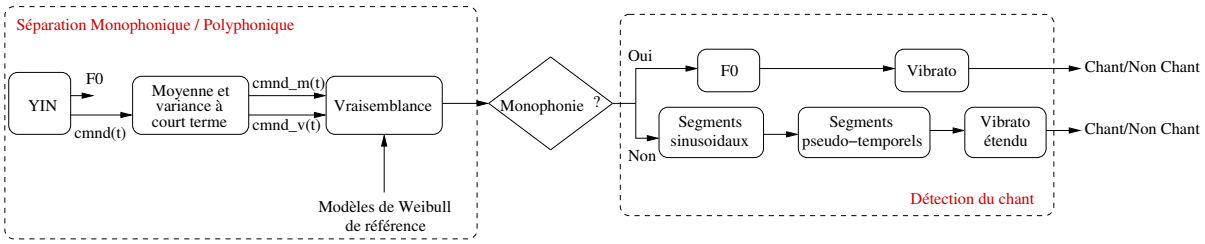


FIG. 4.5 – Schéma général du système de détection du chant.

4.4.3.1 La détection en contexte monophonique

En contexte monophonique, l'estimateur de fréquence fondamentale YIN [dCK02] (partie 3.3.1) est plus performant que l'estimateur de segments sinusoïdaux. En effet, quand la fréquence fondamentale existe, les erreurs faites par l'estimateur YIN sont principalement de donner des multiples (les harmoniques) de la fréquence fondamentale. Or, s'il y a du vibrato sur la fréquence fondamentale, il y en a aussi sur ses harmoniques. *A contrario*, l'estimateur de segments sinusoïdaux trouve effectivement la fréquence fondamentale et ses harmoniques, mais il trouve également des maxima « parasites ». Nous décidons de nous baser sur la fréquence fondamentale estimée par la méthode YIN pour chercher le vibrato.

La présence de vibrato est toujours caractérisée par la présence d'un maximum entre 4 et 8 Hz dans la transformée de Fourier du suivi de fréquence (ici la fréquence fondamentale). Cependant, les ruptures brutales dans la courbe de la fréquence fondamentale (dues aux changements de notes) « bruitent » la transformée de Fourier. Il n'est donc pas possible de segmenter arbitrairement le signal en segments de 1 seconde et de faire la transformée de Fourier de la fréquence fondamentale de chaque segment.

Nous proposons comme alternative de segmenter temporellement la fréquence fondamentale estimée en « notes », en utilisant une méthode proche des « Note Like Unit » proposée par Ohishi *et al.* [OGIT05]. Le but est d'aboutir à des segments dont la définition théorique corresponde à celle de la note de musique. Une transition entre deux notes est trouvée lorsque :

- Soit la fréquence fondamentale fait un saut de plus d'un demi-ton : il y a changement de note.
- Soit il n'y a pas de fréquence fondamentale avant (ou après) la rupture : on se situe au début ou à la fin d'une phrase musicale.

Dès lors, le vibrato est recherché sur chaque note.

La décision finale est prise chaque seconde de la façon suivante : il y a du chant s'il y a du vibrato sur au moins 10% des trames, seuil que nous avons déterminé expérimentalement (voir l'expérience décrite dans la partie 4.5.3).

4.4.3.2 La détection en contexte polyphonique

En contexte polyphonique, nous ne pouvons pas nous baser sur l'estimateur YIN. L'idéal serait d'avoir un estimateur multipitch, capable d'extraire un nombre indéterminé de fréquences fondamentales simultanées. Des travaux sont en cours sur le sujet, mais aucun n'est assez abouti pour nous permettre d'analyser chaque fréquence fondamentale, d'en faire le suivi, et détecter la présence de vibrato.

Nous cherchons ainsi la présence de vibrato sur les segments sinusoïdaux en utilisant le paramètre *vibr* du vibrato étendu que nous avons présenté dans la partie 4.3.3.

La décision finale est prise là aussi chaque seconde, en tenant compte de la nouvelle définition du chant : il y a du chant si, sur au moins une trame, $vibr > 0,15$. Cette fois encore, le seuil a été déterminé expérimentalement.

Nous remarquons que le seuil a changé par rapport à la méthode précédente (il était de 0.08 dans le système primaire [LAOP07]). En effet, d'une part, le fait de séparer *a priori* les sons monophoniques des sons polyphoniques nous permet d'avoir un seuil réellement adapté pour la détection du chant en contexte polyphonique. D'autre part, le fait de ne plus se baser sur une valeur moyennée sur 1 seconde nous a mené à choisir un seuil différent (voir l'expérience décrite dans la partie 4.5.3).

4.5 Expériences

Dans cette partie, nous présentons les diverses expériences que nous avons menées pour tester notre méthode.

Tout d'abord, nous présentons une expérience menée avec un système de base : la paramétrisation est faite avec des MFCC, la modélisation par des GMM. Ensuite, nous

présentons les résultats obtenus par notre système primaire : la paramétrisation est réalisée à l'aide du vibrato étendu, la décision est prise par seuillage. Enfin, nous présentons les résultats obtenus en introduisant le pré-traitement pour distinguer les monophonies des polyphonies, et en tenant compte de notre nouvelle définition du chant.

Le corpus utilisé pour les tests est le même que celui utilisé pour la séparation monophonie / polyphonie (voir partie 3.5.1), avec la même séparation des données entre apprentissage et test. Nous avons ainsi 75 secondes d'apprentissage pour la classe « Chant », et 50 secondes pour la classe « Non Chant ». Ici encore, nous donnons le taux global d'erreur, accompagné de la matrice de confusion.

4.5.1 Système de base

Nous prenons comme système de base, auquel nous nous comparons, le schéma généralement utilisé en classification automatique des données audio. La paramétrisation est réalisée à l'aide de MFCC, et la modélisation avec des GMM. D'après les autres études menées avec ces paramètres, cette méthode a des résultats proches de l'état de l'art [RH07].

Tout comme dans la partie 3.6.2.1, nous testons différentes configurations pour la paramétrisation et pour la modélisation. Pour la paramétrisation, nous avons : soit l'Énergie et 12 coefficients MFCC, soit l'Énergie et 12 coefficients MFCC, auxquels nous ajoutons leurs dérivées. Pour la modélisation, nous faisons varier le nombre de composantes pour les GMM de 1 à 256, avec des matrices de covariance diagonales. Enfin, pour le GMM à une composante, nous testons la configuration dans laquelle la matrice de covariance est pleine.

Pendant la phase d'apprentissage, nous estimons les paramètres des GMM. La prise de décision est prise sur une seconde, par maximum de vraisemblance.

Le tableau 4.1 présente le taux global d'erreur obtenu pour chaque configuration. Pour la détection du chant, la meilleure configuration est obtenue en prenant l'Énergie, 12 coefficients MFCC, et leurs dérivées, et un GMM à 64 composantes. Le taux global d'erreur est de **28,7 %**, la matrice de confusion est présentée dans le tableau 4.2.

Tab. 4.1 – Taux d'erreur pour les différentes configurations testées (en %).

Nb de GMM	1	2	4	8	16	32	64	128	256	1(MP)[39]
E+12 MFCC	39,1	37,9	67,1	66,3	63,4	65,1	39,1	36,6	37,8	61,0
E+12 MFCC+Δ	36,7	37,1	28,9	29,7	29,0	29,8	**28,7**	29,3	30,6	36,6

Nos résultats sont compatibles avec ceux obtenus dans d'autres travaux [RH07, LGD07], qui testent leurs algorithmes sur de la musique Pop. Ils obtiennent, sans post-traitement, les performances suivantes : 25 % d'erreur et une F-mesure de 72,7 %.

TAB. 4.2 – Matrice de confusion obtenue avec la meilleure configuration du système de base.

	Chant	Non chant
Chant	**76,1%**	24,4 %
Non chant	39,7 %	**60,3 %**

De ces résultats, nous pouvons tirer plusieurs conclusions. Tout d'abord, l'utilisation des dérivées dans la paramétrisation semble absolument nécessaire dans cette approche. En effet, les ajouter améliore systématiquement les résultats.

4.5.2 Système primaire : pas de segmentation monophonie / polyphonie

Avec la système primaire [LAOP07] (tel que présenté dans la partie 4.4.1, sans segmentation préalable monophonie / polyphonie).

Pendant la phase d'apprentissage, nous avons un seuil à régler. La valeur optimale est 0,08. La décision est prise chaque seconde par seuillage de la valeur de *vibr* moyennée sur une seconde.

Nous avons un taux d'erreur de **29,7 %**, la matrice de confusion pour cette configuration est présentée dans le tableau 4.3.

TAB. 4.3 – Matrice de confusion obtenue avec le système primaire, sans séparation monophonie / polyphonie.

	Chant	Non chant
Chant	**70,3%**	29,7 %
Non chant	29,6 %	**70,4 %**

Notre système primaire a des performances globalement comparables à celles du système de base : le taux global d'erreur est semblable. Notre approche a un taux de non-détection plus élevé, mais un taux de fausse alarme plus faible. Il nous semble intéressant de noter que notre méthode n'utilise pourtant qu'un seul paramètre, le vibrato étendu, et une règle de décision très simple, puisqu'il s'agit d'une comparaison à un seuil.

Les erreurs de non-détection sont dues à du chant trop faible par rapport aux instruments présents ou à du chant présent pendant une durée trop courte. Dans le premier cas, la valeur du vibrato étendu étant faible, sa moyenne sur une seconde l'est aussi. Dans le

[39] MP : Matrice pleine

second cas, certaines valeurs du vibrato étendu sont bien supérieures au seuil de décision, mais en trop faible nombre pour que la moyenne sur une seconde le soit également.

Les fausses alarmes sont dues à des instrumentistes qui produisent volontairement un vibrato comme effet de style.

4.5.3 Utilisation de la segmentation monophonie / polyphonie

Afin d'évaluer correctement l'apport du pré-traitement lié à la segmentation monophonie / polyphonie, nous avons expérimenté deux situations :

- Le pré-traitement est réalisé manuellement, aucune erreur ne sera imputable au pré-traitement, dont l'apport doit être maximum.
- Le pré-traitement est automatique ; le système global est évalué.

4.5.3.1 Avec une segmentation monophonie / polyphonie manuelle

Le système de détection du chant est couplé avec une segmentation monophonie / polyphonie manuelle. Nous considérons que nous avons là la limite supérieure des performances que nous pouvons atteindre avec cette seule méthode.

Pendant la phase d'apprentissage, nous avons deux seuils à régler, un pour le contexte monophonique, l'autre pour le contexte polyphonique.

Les résultats de cette expérience sont présentés dans les tableaux 4.4, 4.5 et 4.6. Le taux d'erreur global est maintenant de **21,7%**.

Tab. 4.4 – Matrice de confusion obtenue dans le cas monophonique, avec une segmentation monophonie / polyphonie manuelle.

	Chant	Non chant
Chanteur solo	**83 %**	17 %
Instrument solo	20 %	**80 %**

Tab. 4.5 – Matrice de confusion obtenue dans le cas polyphonique, avec une segmentation monophonie / polyphonie manuelle.

	Chant	Non chant
Chanteurs (et instruments)	**66 %**	34 %
Plusieurs instruments	16 %	**84 %**

Nous remarquons tout d'abord que le fait d'utiliser la connaissance *a priori* sur le caractère monophonique ou polyphonique de la musique améliore très nettement les performances pour la détection du chant. Le taux d'erreur global a diminué de 8 %.

TAB. 4.6 – Matrice de confusion globale obtenue avec une segmentation monophonie / polyphonie manuelle.

	Chant	Non chant
Chant	**74 %**	26 %
Non chant	18 %	**82 %**

La détection du chant en contexte polyphonique est plus difficile qu'en contexte monophonique. Les non-détections sont dans ce cas dues au fait que la voix est faible par rapport aux instruments présents.

En contexte monophonique, les fausses alarmes sont dues à des instruments à vent sur lesquels le musicien produit volontairement un vibrato. Encore utilisé aujourd'hui dans certains styles musicaux, notamment chez les violonistes, c'est un effet d'ornementation qui fut très en vogue à certaines époques dans la musique classique. Sur un violon, où il est produit en faisant vibrer le doigt qui touche la corde, le vibrato a un rythme assez élevé (supérieur à 10 Hz). Les violons ne sont donc pas classés comme chanteurs. Par contre, sur les instruments à vent, il est produit en faisant varier la puissance du souffle, ce qui se fait à un rythme proche de 5 Hz. Ainsi, quand un flûtiste introduit un effet de vibrato, le son résultant sera classé comme du chant.

Le vibrato ne pourra pas seul venir à bout de ces erreurs, la solution sera d'utiliser d'autres paramètres.

4.5.3.2 Avec une segmentation monophonie / polyphonie automatique

Dans un second temps, nous nous proposons de tester l'ensemble du système. La séparation monophonie / polyphonie est faite en utilisant le système que nous avons développé (voir chapitre 3). En n'utilisant que des modules automatiques pour le prétraitement (5 modèles de Weibull pour la séparation monophonie / polyphonie) et pour la détection du chant, les résultats restent malgré tout encourageants puisque le taux d'erreur global est de **25 %**. Les résultats par classe et par sous-classe sont présentés dans les tableaux 4.7 et 4.8.

TAB. 4.7 – Matrice de confusion globale obtenue avec une segmentation monophonie / polyphonie automatique.

	Chant	Non chant
Chant	**72 %**	28 %
Non chant	21 %	**79 %**

TAB. 4.8 – Résultats détaillés obtenus avec une segmentation séparation monophonie / polyphonie automatique.

	Chant	Non chant
Chanteur solo	**79 %**	21 %
Instrument solo	26 %	**74 %**
Chanteurs (et instruments)	**65 %**	35 %
Plusieurs instruments	18 %	**82 %**

Nous remarquons que la détection du chant en contexte polyphonique est toujours plus difficile qu'en contexte monophonique. Cependant, nous remarquons que, par rapport aux résultats obtenus avec une segmentation monophonie / polyphonie manuelle, les résultats sont surtout dégradés en contexte monophonique (on passe de 83 % et 80 % de bonnes détections à 79 % et 74 %) alors qu'ils sont quasiment équivalents en contexte polyphonique (on passe de 66 % et 84 % de bonne détection à 65 % et 82 %). Ceci est en adéquation avec les résultats que donne la méthode de séparation monophonie / polyphonie : il y a plus de monophonies classées comme polyphonies (11,3 %) que l'inverse (4,8 %) (voir tableau 3.15).

Si un extrait monophonique est détecté comme polyphonique, c'est que la fréquence fondamentale n'est pas bien définie, probablement parce qu'il y a du bruit ou que le son n'est pas clair. Les outils utilisés pour détecter la présence de chant sont alors les segments sinusoïdaux et le vibrato étendu. Comme la fréquence fondamentale n'est pas bien définie, nous détectons de mauvais segments sinusoïdaux : ils ne correspondent ni à la fréquence fondamentale ni à ses harmoniques. Ils interviennent cependant dans le calcul du vibrato étendu et le font pencher parfois dans le mauvais sens.

La solution pour contrer ces erreurs peut venir soit de l'amélioration de la segmentation monophonique / polyphonique, soit d'une amélioration de la détection du chant en contexte polyphonique.

Les autres erreurs sont dues aux mêmes causes que dans l'expérience précédente (chant masqué par des instruments, introduction volontaire de vibrato...).

4.6 Conclusion

Nous avons présenté une méthode de détection du chant, basée sur la détection de vibrato. Afin de pouvoir utiliser cette notion en contexte polyphonique, nous l'avons étendue en proposant un nouveau paramètre : le vibrato étendu. L'utilisation de ce seul paramètre pour la classification donne des résultats comparables à l'état de l'art : le taux d'erreur est de 29,7 %.

Pour améliorer cette détection, nous avons proposé d'intégrer comme pré-traitement la séparation entre les sons monophoniques et les sons polyphoniques (voir le chapitre 3). Ceci nous a permis d'adapter plus précisément notre modèle, tant pour le cadre polyphonique que pour le cadre monophonique, menant à une nette amélioration des résultats.

La prise en compte de la réalité du chant, qui tient compte des différents types de pauses que le chanteur peut faire, nous a amenés à changer notre stratégie de détection du chant. Ces deux améliorations permettent d'obtenir un taux d'erreur de 25 %.

Les erreurs que nous avons sont dues principalement à de la voix recouverte par d'autres instruments, et à du vibrato volontairement produit par des musiciens. Dans les deux cas, le vibrato seul (ou le vibrato étendu) ne suffit pas, la solution nous semble être d'utiliser des paramètres complémentaires.

Un post-traitement approprié pourrait certainement réduire les erreurs. Ce post-traitement devra être déterminé en fonction de l'application finale.

Chapitre 5

Conclusion et perspectives

Sommaire

5.1 Conclusion

Au cours de ce travail de thèse, nous avons proposé une méthode de détection du chant afin d'affiner la décomposition du flux audio en ses composantes primaires que sont la parole et la musique. Cette détection, afin d'être la plus efficace possible, s'appuie sur une distinction préalable de la musique en zones monophoniques et polyphoniques.

5.1.1 La distinction Monophonie / Polyphonie

La méthode pour distinguer les sons monophoniques des sons polyphoniques est fondée sur l'analyse du comportement d'une mesure de confiance issue de l'algorithme YIN au travers de sa moyenne et variance à court terme. Pour modéliser la répartition bivariée de ce vecteur d'observation, nous avons utilisé des lois de Weibull bivariées, qui sont plus appropriées que les Gaussiennes ou Mélanges de Gaussiennes habituellement utilisées dans l'analyse du son. L'apprentisssage de ce type de loi a nécessité de définir une méthode d'estimation du facteur de corrélation, basée sur la méthode des moments.

L'expérimentation a montré qu'un ensemble d'apprentissage relativement restreint, est suffisant : il contient environ 2 minutes de signal composé de données très variées. Sur un ensemble de test d'une durée d'environ 18 minutes, des résultats très intéressants ont

été obtenus, puisque le taux global d'erreur est de 6,3 % d'erreur, contre 19,2 % pour l'approche classique (MFCC modélisés par des GMM).

Nous concevons l'utilisation de cet algorithme comme un pré-traitement, notamment pour la détection du chant. En fonction de l'application finale, et du corpus étudié, on pourrait appliquer un post-traitement approprié, afin d'en améliorer les résultats. L'unité de décision étant présentement la seconde, il est certain qu'un lissage est immédiatement envisageable !

Compte tenu des corpora utilisés, les modèles de Weibull bivariés ainsi estimés sont directement utilisables sur de nouvelles données. Cet outil est simple et suffisamment robuste pour les traiter sans qu'il soit nécessaire de ré-annoter de nouveaux exemples – très coûteux, ni de ré-apprendre les modèles.

5.1.2 La détection du chant

La détection du chant se base sur la détection de vibrato, un paramètre qui est défini à partir de l'analyse de la fréquence fondamentale. Afin de pouvoir l'utiliser en contexte polyphonique, nous avons, à l'aide de deux segmentations, une segmentation sinusoïdale et une segmentation pseudo-temporelle, introduit un nouveau paramètre : le *vibrato étendu*, une quantification du vibrato sur les principales harmoniques.

Les expérimentations montrent qu'utilisé brutalement sans prétraitement monophonie/polyphonie, les résultats obtenus par la méthode sont comparables à l'état de l'art, avec un taux d'erreur global de 29,7 %.

Lorsque la distinction entre les sons monophoniques et les sons polyphoniques est mise en œuvre comme prétraitement, nous avons adapté et optimisé notre détection du chant pour chaque contexte. Le taux global d'erreur s'abaisse à 25 %.

Notons que, dans le cadre d'une application bien définie, il est tout à fait possible d'envisager un post-traitement approprié.

L'apprentissage est fait sur un corpus suffisamment hétérogène pour penser que les seuils de décision sont robustes. La méthode peut donc être considérée comme « sans apprentissage », dans de futures mises en œuvre.

5.1.3 Bilan sur la structuration d'un document

Avec cet ultime détection du chant, nous sommes en mesure de structurer le flux audio en ses différentes composantes primaires que sont la parole, la musique, le chant et les jingles :

Les jingles L'outil de détection a été développé par Julien Pinquier et Régine André-Obrecht [PAO04]. Un jingle est caractérisé par sa signature spectrale qui est retrouvée par simple calcul de distance sur le flux audio.

La parole et la musique Les outils de détection de la parole et de musique ont été développés par Julien Pinquier et Régine André-Obrecht [PRAO03]. Ils se basent sur l'analyse de quatre paramètres : la modulation de l'énergie à 4 Hz et la modulation de l'entropie pour la parole ; la durée moyenne et le nombre de segments par secondes, segments issus d'une segmentation en zones pseudo-stationnaires [AO88] pour la musique. Un simple seuillage sur les paramètres conduit à la détection.

La stratégie employée afin d'obtenir une structuration primaire est la suivante :

1. Le premier niveau de segmentation du document se fait à l'aide des jingles. Plusieurs approches sont possibles pour le choix des jingles que nous allons détecter :
   - Nous pouvons prendre tous les jingles du média considéré. Ceci permet d'obtenir une segmentation fine, et surtout d'avoir facilement le titre de l'émission. Cependant, nous avons un risque plus élevé de fausse alarme.
   - Nous pouvons sélectionner quelques jingles selon leur importance en terme de structuration du media considéré. Il s'agit pour une radio ou une chaine télévisuelle donnée des jingles des journaux d'information, de la météo, et des plages de publicité ; ils séparent systématiquement des émissions, et il existe toujours au moins l'un de ces jingles entre deux émissions successives.

   C'est ce dernier choix que nous privilégions. Nous appellons dès lors *émission* l'intervalle de temps compris entre deux jingles successifs.

2. Au sein d'une émission, nous analysons les plages de parole et de musique détectées. La simple analyse de la durée de ces plages, de leur nombre, et de leur disposition temporelle nous permet déjà de caractériser en partie l'émission. Nous séparons ainsi plusieurs grands types d'émissions :

   Les plages publicitaires Les plages publicitaires sont délimitées par des jingles spécifiques, leur durée est courte : légalement, une chaine ne peut diffuser plus de 6 minutes de publicité par heure en moyenne. Les publicités contiennent de la parole, souvent superposée à de la musique.

   Les émissions musicales Au sein de ces émissions, il y a de nombreuses plages de musique, la durée totale de la musique est élevée. Les plages de musique sont séparées par des zones de parole de durée moyenne.

   Les émissions d'information Journaux télévisés ou radiophoniques, les journaux d'information sont reconnaissables à leur jingle, et au fait que les plages de musiques sont soit inexistantes, soit en faible nombre et de courte durée. La principale composante de ces émissions est la parole, avec parfois des zones de bruit.

Les émissions culturelles ou de divertissement À la radio, ces émissions sont souvent rythmées par des pauses musicales. Celles-ci sont peu nombreuses, et de durée moyenne. Entre elles, il y a de longues zones de parole.

Les films À la télévision, les films sont caractérisés par une durée de musique élevée, une grande durée de la musique de fond (Parole + Musique), et de nombreuses zones de bruit.

3. Une dernière précision est obtenue en détaillant le contenu des plages de musique en notant la présence ou non de chant.

5.1.4 Application sur une émission

Nous illustrons cette démarche sur un exemple de structuration d'une journée de radio, la journée du 19 février 2009, entre 00h00 et 24h00.

Sur toutes les figures, le même code de couleur est utilisé :
- les plages bleues (première ligne) correspondent aux jingles,
- les zones oranges (deuxième ligne) représentent les zones de musique,
- les zones vertes (troisième ligne) sont les zones de chant.

Étape 1 : Structuration d'une journée – Détermination du type d'émission Dans un premier temps, nous nous basons sur l'analyse des jingles, de la parole et de la musique pour segmenter le flux en émissions.

La segmentation en émissions obtenue est présentée sur la figure 5.1. Nous remarquons qu'il y a des jingles toutes les heures. Entre 05h00 et 09h00, nous sommes en présence d'une plage matinale d'information, dans laquelle le journal est répété toutes les demi-heures, avec donc des jingles toutes les demi-heures. Enfin, le jingle de 23h00 n'est pas détecté, mais une analyse de la régularité des jingles nous permet d'inférer qu'il s'agit en réalité d'une non-détection.

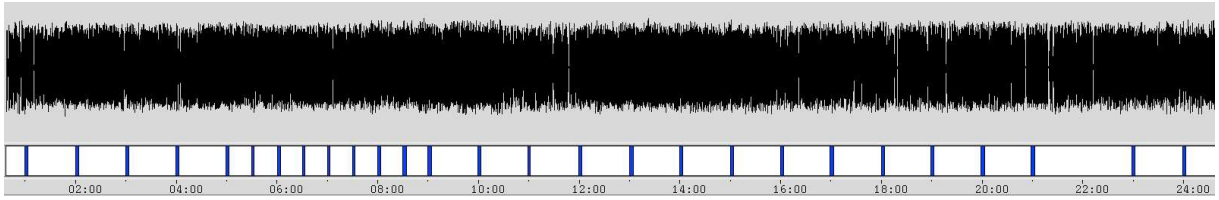


FIG. 5.1 – France Inter – Structuration de la journée par les jingles

Pour chacune des émissions, l'analyse du nombre de plages de musique et de leur durée nous permet de leur attribuer un type. Trois exemples sont présentés ci-dessous :

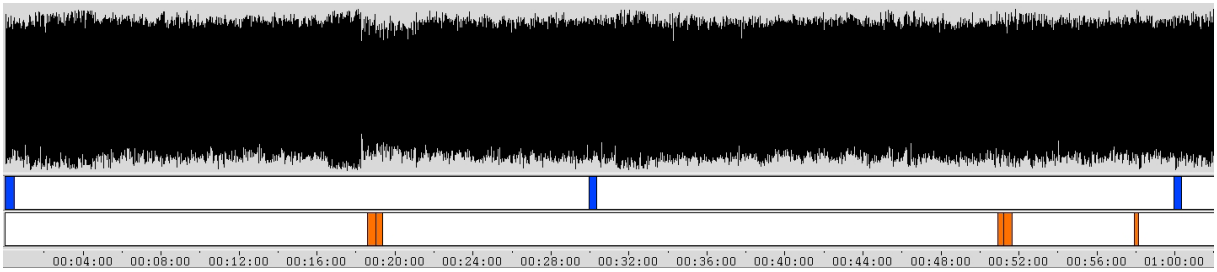


FIG. 5.2 – France Inter – 06h00-07h00 : Plage d'information

– La plage horaire 06h00-07h00 (voir figure 5.2) correspond à une plage d'information. Il n'y a quasiment pas de musique, les seules courtes plages détectées sont des publicités / auto-promotions, et une fausse-alarme.
– La plage horaire 11h00-12h00 (voir figure 5.3) est une émission de divertissement. On y retrouve principalement deux plages de musique de plusieurs minutes chacune, qui correspondent à deux pauses musicales. Les autres plages (courtes) sont des fausses-alarmes.

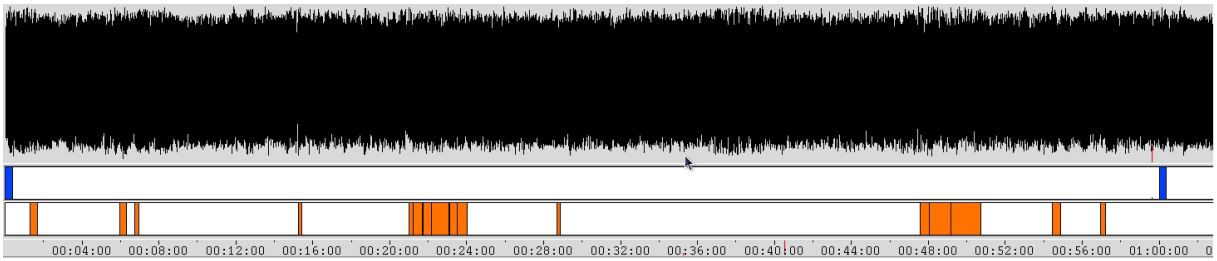


FIG. 5.3 – France Inter – 11h00-12h00 : Émission de divertissement

– La plage horaire 16h00-17h00 (voir figure 5.4) est une émission musicale. De nombreux extraits de musique de plusieurs minutes chacun sont présents. La durée totale de musique est supérieure à 50 % du temps de l'émission.

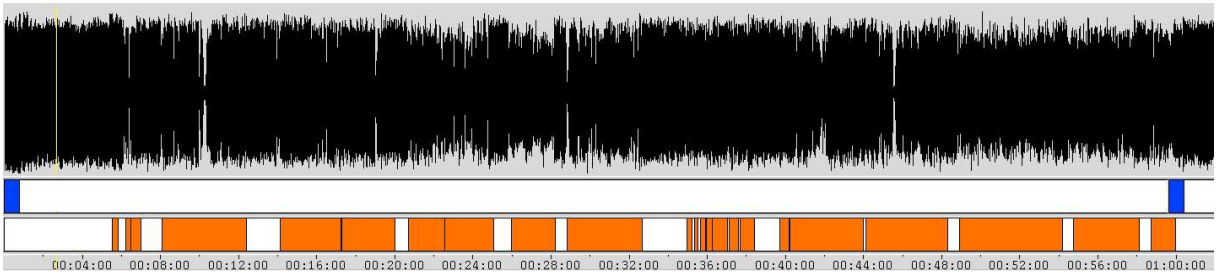


FIG. 5.4 – France Inter – 16h00-17h00 : Émission musicale

Étape 2 : Émissions musicales – Précisions sur la nature de la musique Dans un deuxième temps, nous précisons le contenu des émisions musicales. Celles-ci sont au

nombre de deux dans la journée : une émission de musique instrumentale dans l'après-midi, et un émission de musique de variétés en début de nuit. Pour chacune de ces émissions, nous analysons les plages musicales en terme de présence de chant. Les résultats sont présentés sur les figures 5.5 et 5.6.

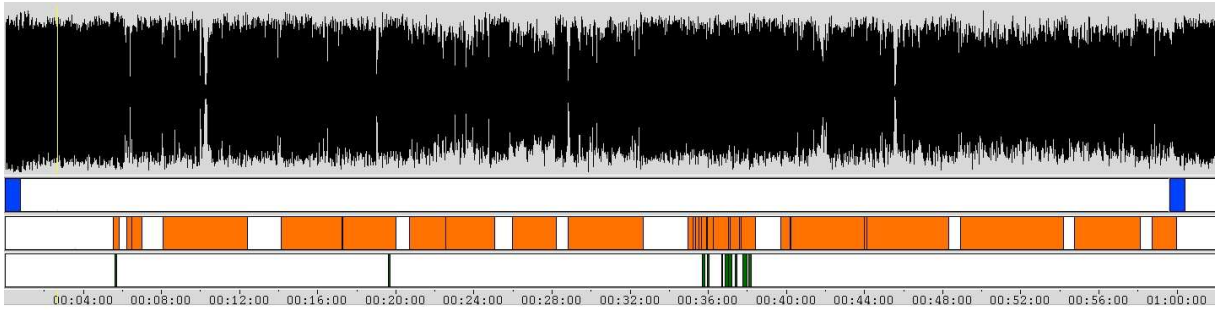


FIG. 5.5 – France Inter – 16h00-17h00 : Emission musicale, musique purement instrumentale

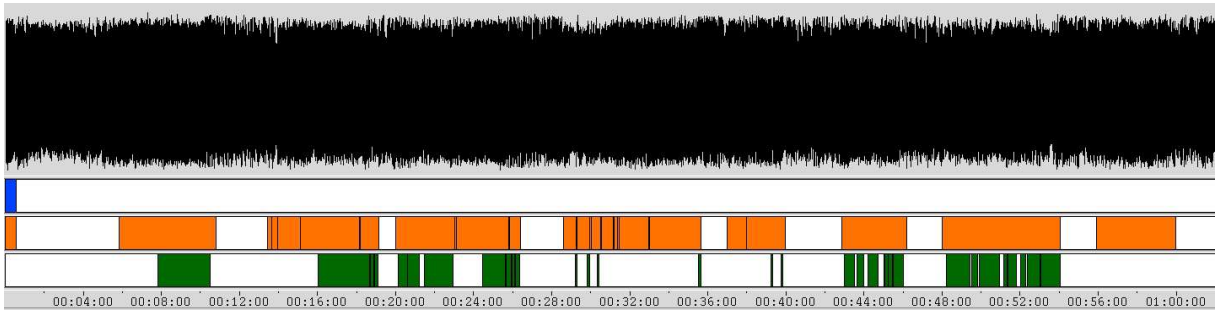


FIG. 5.6 – France Inter – 22h00-23h00 : Emission de variétés

5.2 Perspectives

Comme en témoigne le paragraphe précédent, l'indexation élémentaire du flux audio peut largement être complétée. Les méthodes scientifiques doivent être complétées et les applications élargies.

5.2.1 Sur les méthodes

Une tentative de séparation Parole seule / Parole simultanée
La méthode développée pour déterminer si une ou plusieurs sources harmoniques sont présentes (polyphonie/monophonie) est efficacement appliquée au contexte musical. Il nous semble qu'une méthode semblable pourrait être développée pour la détection des zones où, dans une conversation, plusieurs personnes parlent en même temps. Il s'agit là

aussi de différentier, d'un coté « une personne » (la classe « Parole Seule »), d'un autre, « plusieurs personnes » (la classe « Parole Simultanée »).

Nos premières expériences montrent qu'effectivement les répartitions bivariées de nos paramètres pour ces deux classes sont différentes. Cependant, la présence de zones non voisées, dues à la présence de consonnes dans la voix, a tendance à rendre ces deux distributions trop semblables pour que notre méthode soit applicable en l'état. Une solution consiste à ne considérer que les zones voisées qu'il faut alors délimiter de façon précise. Cette première extension est en cours d'étude.

Vers une fusion de plusieurs méthodes pour la détection du chant ?
En analysant de près les résultats obtenus par notre méthode et ceux d'une méthode « état de l'art » (MFCC modélisés par des GMM), il nous semble que de meilleurs résultats pourraient être obtenus en les fusionnant. Les deux méthodes donnent des résultats équivalents mais les erreurs sont complémentaires : notre méthode a un taux de non-détection plus élevé, mais un taux de fausse-alarme plus faible que la méthode « état de l'art ». Une telle stratégie de fusion implique de pouvoir attribuer des scores de confiance pour chaque méthode, de choisir un modèle de fusion qui peut impliquer une normalisation des scores entre méthodes.

5.2.2 Sur la description des contenus audio par leur contenu musical

De l'importance des jingles Le premier point qu'il nous semble important de souligner est le rôle primordial des jingles. Comme le souligne Anne-Marie Gustave [Gus08], les radios n'hésitent pas à investir, et à faire appel à de grands compositeurs de musiques de film pour créer leur « habillage sonore ».

De fait, dans l'indexation audio, ils sont utiles à plusieurs points de vue. Tout d'abord, ils déterminent les limites des émissions, et les identifient. Mais ils donnent aussi des indices sur le genre d'émission : une émission de divertissement n'a pas le même générique qu'une émission de musique ou d'information.

Enfin, les jingles peuvent être structurants au sein d'une émission : ils rythment les émissions, séparent deux sujets dans les journaux d'information, ou encore annoncent des séquences récurrentes.

Dans la mesure où notre méthode de détection de jingles permet de détecter quasiment l'ensemble des jingles d'une radio (d'une chaîne de télévision) à partir de quelques jingles de base de cette chaîne, ne serait-il pas intéressant de favoriser cette sur-détection et d'envisager ensuite leur caractérisation en genre ?

Quelques réflexions sur la caractérisation des extraits musicaux Les outils de détection de la musique sont utiles à plusieurs titres. La seule information donnée par la

durée de chaque extrait musical, ou encore par le nombre d'extraits musicaux ou chantés diffusés au cours d'une émission sont déjà de bons indices pour la description d'une émission, comme nous l'avons proposé ci-dessus. Il serait intéressant de compléter cette caractérisation en faisant appel aux nombreux outils de classification ou de similarité.

De la musique de fond, cas des films L'analyse de la musique de fond est, nous semble-t-il, absolument primordiale, notamment pour l'analyse de films.

La simple détection de la présence de musique est déjà très porteuse d'information. Ce problème est particulièrement difficile dans les films : la musique est souvent recouverte de parole, mais elle est superposée aussi à des bruits divers (cris, bruits de voiture...). Les travaux sur la détection de la musique de fond sont relativement avancés. En revanche très peu existent sur sa description. Une étude a été menée par Abe et Nishiguchi [AN02] pour rechercher une musique connue qui serait utilisée comme musique de fond.

Pourtant, la connaissance d'informations telles que le rythme, le genre, ou la tonalité de la musique de fond serait très utile pour décrire la scène : schématiquement, nous pouvons imaginer qu'une scène d'action aura un rythme rapide, une scène d'adieu une tonalité mineure...

La caractérisation de la musique de fond nous semble être un point clé pour la caractérisation des documents audio par la musique, point qu'il sera important d'approfondir dans un avenir proche. Il n'est pas sûr que les outils développés pour les extraits musicaux suffisent pour servir de base à ces recherches.

Annexe A

Mesures de performances

A.1 Le Taux d'Erreur Global

Le Taux d'Erreur Global (Global Error Rate ou GER) en anglais est défini de la manière suivante :

$$Err = \frac{Nombre\ d'échantillons\ mal\ classifiés}{Nombre\ total\ d'échantillons} \tag{A.1}$$

Le taux d'erreur global indique la proportion d'échantillons mal classifiés, sans distinguer les types d'erreurs (A classé comme B ou B classé comme A).

A.2 La Matrice de Confusion

Pour un problème à n classes, la matrice de confusion est un tableau $n*n$, qui résume les erreurs faites pendant la classification, en précisant, pour chaque classe, quelle proportion d'échantillons a été classée dans chacune des classes. Un exemple est donné, pour trois classes dans le tableau A.1.

Tab. A.1 – Matrice de confusion - Exemple.

	Classe 1	Classe 2	Classe 3
Classe 1	**95,6 %**	2.5 %	1.9 %
Classe 2	0 %	**75.2 %**	24.8 %
Classe 3	2.5 %	15.4 %	**82.1 %**

Dans cet exemple la classe 1 est bien reconnue, alors que les classes 2 et 3 sont relativement plus difficiles à distinguer. Cette présentation des résultats est certes moins synthétique que le taux d'erreur global, mais permet une analyse plus fine des erreurs – et permet donc de déterminer des pistes de travail pour l'amélioration des algorithmes.

A.3 La Précision et le Rappel

Ces mesures sont utilisées pour des tâches de type extraction d'information : il s'agit de problèmes posés sous la forme Classe/Non classe, et non sous la forme Classe1/Classe2. Elles sont définies à partir des notions de « Vrai Négatif » (VN), « Faux Négatif » (FN), « Vrai Positif » (VP) et « Faux Positif » (FP), résumées dans le tableau A.2.

TAB. A.2 – Tableau résumé des notions de « Vrai Négatif », « Faux Négatif », « Vrai Positif »et « Faux Positif ».

		Données prédites	
		Vrai	Faux
Vérité Terrain	Vrai	Vrai Positif	Faux Négatif
	Faux	Faux Positif	Vrai Négatif

La Précision (precision en anglais) est définie de la manière suivante :

$$R = \frac{Nombre\ d'échantillons\ pertinents\ trouvés}{Nombre\ d'échantillons\ trouvés} = \frac{VP}{VP + FP} \tag{A.2}$$

La précision mesure donc la proportion de documents pertinents parmi ceux trouvés. Une précision de 100 % signifie que tous les documents trouvés sont pertients.

Le Rappel (recall en anglais) est défini de la manière suivante :

$$P = \frac{Nombre\ d'échantillons\ pertinents\ trouvés}{Nombre\ d'échantillons\ pertinents\ dans\ la\ base} = \frac{VP}{VP + FN} \tag{A.3}$$

Le rappel mesure donc la proportion de documents trouvés parmi ceux cherchés. Un rappel de 100 % signifie qu'on a trouvé tous les documents cherchés.

A.4 La F-Mesure

La F-Mesure est définie à partir de la précision et du rappel de la manière suivante :

$$F = \frac{2PR}{P + R} \tag{A.4}$$

avec P la précision et R le rappel.

La F-mesure est souvent vue comme une altenative au taux d'erreur global. Une utilisation intéressante est le cas où deux classes sont présentes mais très déséquilibrées : par exemple $Card(Classe1) \gg Card(Classe2)$. Dans ce cas, répondre toujours "Classe1" peut être très intéressant si on considère le taux d'erreur global. En effet, on a alors $err = \frac{Card(Classe2)}{Card(Classe1)+Card(Classe2)}$, qui est très petit. Ce genre d'approche donne en revanche une F-mesure nulle. En effet, on a $R = 0$ pour la Classe2.

A.5 L'Accuracy

L'Accuracy est, de manière générale, définie comme le pourcentage d'objets bien reconnus par rapport au nombre d'objets à reconnaitre. En reconnaissance de la parole, c'est par exemple le pourcentage de mots bien reconnus par rapport au nombre de mots attendus.

Dans le cas d'un problème Classe/Non classe, elle est définie de la manière suivante :

$$Acc = \frac{VP + VN}{VP + FP + FN + VN} \tag{A.5}$$

Dans le cas à deux classes, on a alors la relation suivante entre l'accuracy Acc et le taux global d'erreur Err :

$$Acc = 1 - Err \tag{A.6}$$

Annexe B

Test de Kolmogorov

B.1 Descriptif du test

Le test de Kolmogorov est un test d'hypothèse qui permet de décider si une variable aléatoire X suit une loi de fonction de répartion $F(x)$ connue. On teste l'hypothèse H_0 contre l'hypothèse H_1 :

- H_0 : La variable aléatoire suit loi de une fonction de répartion $F(x)$,
- H_1 : La variable aléatoire ne suit pas une loi de fonction de répartion $F(x)$.

Le test est le suivant :

1. Estimation de la fonction de répartition empirique $\hat{F}(x)$ à partir de K observations x_k de la variable aléatoire X. Cette estimation est réalisée avec l'histogramme cumulé, calculé en prenant K classes.
2. Recherche de l'écart maximum Δ_{max} entre $F(x)$ et $\hat{F}(x)$,
3. Pour un risque de première espèce α donné, on accepte H_0 si $\Delta_{max} < \Delta_{Kolmo}$.

B.2 Table du test de Kolmogorov

La valeur de l'écart maximum théorique Δ_{Kolmo} entre l'histogramme cumulé et la fonction de répartition dépend uniquement du nombre d'échantillons N_c disponitbles et du risque de première espèce α. Pour un nombre d'échantillons inférieur à 100, et pour $\alpha = 5$ % et $\alpha = 1$ %, la table B.1 donne les valeurs de Δ_{Kolmo}[40]

B.3 Cas où $N_c > 100$

Dans le cas où le nombre d'échantillons disponibles est supérieur à 100, les valeurs de Δ_{Kolmo} ne sont pas disponibles.

[40]Des tables existent pour toutes les valeurs de N_c comprises entre 1 et 100, et pour $\alpha = 1$ %, $\alpha = 2$ %, $\alpha = 5$ %, $\alpha = 10$ % et $\alpha = 20$ %. Ces tables sont disponibles en ligne sur Internet.

TAB. B.1 – Test de Kolmogorov : valeur de l'écart maximum théorique Δ_{Kolmo}.

N_c	$\alpha = 5$ %	$\alpha = 1$ %
5	0.5633	0.6685
10	0.4087	0.4864
15	0.3375	0.4042
20	0.2939	0.3524
25	0.2639	0.3165
30	0.2417	0.2898
40	0.2101	0.2521
50	0.1884	0.2260
60	0.1723	0.2067
70	0.1597	0.1917
80	0.1496	0.1795
90	0.1412	
100	0.1340	

Une possibilité est d'utiliser comme valeur approchée $\Delta_{Kolmo} = \frac{1.358}{\sqrt{N_c}}$ pour $\alpha = 5$ % et $\Delta_{Kolmo} = \frac{1.629}{\sqrt{N_c}}$ pour $\alpha = 1$ %. La fonction Matlab *kstest* calcule cette valeur.

Une autre possibilité est d'approximer directement $P(max_x|\hat{F}(x) - F(x)|) > \Delta_{Kolmo}$ de la manière suivante [Vap00].

On sait que :

$$\begin{aligned} \alpha = P[\text{rejeter } H_0 | H_0 \text{ vraie}] =& P(\max_x |\hat{F}(x) - F(x)| > \Delta_{Kolmo}) \\ &\simeq \sum_{k=1}^{k=+\infty} 2(-1)^{k-1} \exp -2k^2 N_c \Delta_{Kolmo}^2 \end{aligned} \tag{B.1}$$

En mesurant $\Delta_{max} = \max_x |\hat{F}(x) - F(x)|$, on en déduit :

$$P_{\Delta_{max}} = P(\max_x |\hat{F}(x) - F(x)| > \Delta_{max}) \simeq \sum_{k=1}^{k=+\infty} 2(-1)^{k-1} \exp -2k^2 N_c \Delta_{max}^2 \tag{B.2}$$

Comme $P_{\Delta_{max}}$ est une fonction décroissante de Δ_{max}, on en déduit que si $P_{\Delta_{max}} > \alpha$, alors $\Delta_{max} < \Delta_{Kolmo}$. Ainsi, le calcul de Δ_{Kolmo} n'est pas nécessaire, on accepte l'hypothèse H_0 si :

$$\boxed{\sum_{k=1}^{k=+\infty} 2(-1)^{k-1} \exp -2k^2 N_c \Delta_{max}^2 > \alpha} \tag{B.3}$$

Annexe C

Détail du corpus

Nous donnons ci-dessous la liste complète des morceaux de musique dont est extrait notre corpus. Dans les tableaux C.1 et C.2 nous précisons, les morceaux utilisés pour l'apprentissage et le test, et indiquons les sous-classes présentes dans les extraits sélectionnés.

- Barbara – Au Bois de Saint-Amand, 1964
- Suzanne Vega – Tom's Diner, 1987
- Steeleye Span – A Calling-On Song, 1970
- Caudin de Sermisy – Languir me fais, Un jour Robin, 1528
- Clément Marolt – La Maitre Pierre, vers 1550
- Mozart – Les Noces de Figaro, 1786, Requiem en ré mineur, 1791
- Monteverdi – L'Orfeo, 1607
- Malicorne – Marion les Roses, 1975
- Leonhard Lechner – Magnificat Primi Toni, 1578
- Ekdahl – Now or Never, 1998,
- Cowboy Junkies – Minning for Gold, 1988
- Georg Philipp Telemann
- Publicité pour le Mémorial de Caen
- Antonín Dvořák – Symphonie du Nouveau-Monde, 1875
- Jacob van Eyck – Der Fluyten Lust-hof, 1644
- Tortoise – vers 1995
- Django Reinhardt – Blues d'Autrefois, 1943
- Duke Ellington – Switch Blade, 1962
- Cali – Elle m'a dit, 2003
- Dominique A – Le Twenty-Two Bar, 1995
- Richard Wagner – Der fliegende Holländer, 1843
- Giuseppe Verdi – Aïda, 1871, Otello, 1887
- Paris Combo – Terrien d'Eau Douce, 1999
- Anonyme – Llibre Vermel de l'Abbaye de Montserrat, XIV^{ième} siècle
- Thomas Morley – Sweet Nymph, vers 1600

- Jerry Lee Lewis – Great Bass of Fire, 1957
- Tri Yann – Kiss the Children for Me Mary, La Jument de Michao, 1976
- Georges Bizet – Carmen, 1875
- Beck – Lazy Flies, 1998
- Elton John – Never Knew Her Name, 1989
- Philippe Katerine – Le Jardin Anglais, 1996
- Jimmy Hendrix – Can You See Me, 1967
- Bob Dylan – Lay Daly Lay, 1969
- Ben Harper – Excuse Me Mister, 1995
- Les Hurlements d'Léo – L'Accordéoniste, 1998
- Radiohead – Optimistic, 2000

Tab. C.1 – Origine des extraits utilisés pour l'apprentissage.

Auteur/Titre	Chant mono	Chant poly	Inst. mono	Inst. poly	Inst.+chant
Barbara	×				
Suzanne Vega	×				
Mozart – Noces	×				
Ekdahl	×				
Cowboy Junkies	×				
Steeleye Span		×			
Sermisy – Languir		×			
Monteverdi		×			
Malicorne		×			
Lechner (homme)		×			
Dvořák			×		
Van Eyck			×		
Tortoise			×		
Django Reinhart			×		
Duke Ellington			×		
Publicité				×	
Anonyme – Llibre				×	
Telemann				×	
Bob Dylan				×	
Ben Harper				×	
Cali – Elle m'a dit					×
Wagner					×
Verdi – Aïda					×
Paris Combo					×
Elton John					×

TAB. C.2 – Origine des extraits utilisés pour le test.

Auteur/Titre	Chant mono	Chant poly	Inst. mono	Inst. poly	Inst.+chant
Alfred Deller	×	×			
Sermisy – Un Jour	×	×			
Jerry Lee Lewis	×	×	×		
Lechner (femme)	×				
Tri Yann – Kiss	×		×		×
Philippe Katerine	×			×	×
Clément Marolt		×			
Verdi – Otello		×			
Bizet		×	×	×	
Telemann			×		
Dominique A			×	×	×
Tri Yann – Michao			×		
Beck				×	×
Jimmy Hendrix				×	
Les Hurlements				×	×
Radiohead				×	
Mozart – Requiem					×

Bibliographie

[AC04] I. Arroabarren and A. Carlosena. Vibrato in Singing Voice : The Link between Source-Filter and Sinusoidal Models. *EURASIP Journal on Applied Signal Processing*, 7 :1007–1020, 2004.

[AC07] I. Arroabarren and A. Carlosena. Voice Production Mechanisms of Vocal Vibrato in Male Singers. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1) :320–332, Jan 2007.

[ADR04] M. Alonso, B. David, and G. Richard. Tempo and beat estimation of musical signals. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, 2004.

[AN02] M. Abe and M. Nishiguchi. Self-Optimized Spectral Correlation Method for Background Music Identification. In *IEEE International Conference on Multimedia and Expo (ICME '02)*, pages 333–336, 2002.

[AO88] R. André-Obrecht. A New Statistical Approach for the Automatic Segmentation of Continuous Speech. *IEEE Transaction on Acoustics, Speech, and Signal Processing*, 36 :29–40, 1988.

[AP03] J.-J. Aucouturier and F. Pachet. Representing Musical Genre : A State of the Art. *Journal of Neaw Music Research*, 32(1) :83–93, 2003.

[Bau72] L. Baum. An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process. *Inequalities*, 3 :1–8, 1972.

[BBP04] J. M. Brück, S. Bres, and D. Pellerin. Construction d'une signature audio pour l'indexation de documents audiovisuels. In *Compression et Représentation des Signaux Audiovisuels (CORESA'2004)*, 2004.

[BCE+06] J. Bergestra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl. Aggregate Features and AdaBoost for Music Classification. *Machine Learning*, 65(2–3) :473–484, 2006.

[BDS06] J. Bello, L. Daudet, and M. Sandler. Automatic Piano Transcription Using Frequency and Time-Domain Information. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6) :2242–2251, 2006.

[BE01] A. Berenzweig and D. Ellis. Locating Singing Voice Segments Within Music Signals. In *Workshop on the applications of signal processing to audio and acoustics*, pages 119–122, 2001.

[BEL02] A. Berenzweig, D. Ellis, and S. Lawrence. Using Voice Segments to Improve Artist Classification of Music. In *Proc. of the 22th AES International Conference*, 2002.

[BF08] B. Bigot and I. Ferrané. From Audio Content Analysis to Conversational Speech Detection and Characterization. In *ACM SIGIR Workshop*, pages 62–65, 2008.

[BGV92] B. Boser, I. Guyon, and V. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *COLT '92 : Proceedings of the fifth annual workshop on Computational Learning Theory*, pages 144–152, 1992.

[BPR+05] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. In *Interspeech - European Conference on Speech Communication and Technology*, 2005.

[BR88] G. Bloothooft and Plomp R. The timbre of sung vowels. *Journal of the Acoustical Society of America*, 84(3) :847–860, 1988.

[Bre94] A. Bregman. *Auditory Scene Analysis : the Perceptual Organization of Sound*. MIT Press, 1994.

[Bro91a] J. Brown. Calculation of a Constant-Q Spectral Transform. *Journal of the Acoustical Society of America*, 89(1) :425–434, 1991.

[Bro91b] J. Brown. Determination of Musical Meter using the Method of Autocorrelation. In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 171–172, 1991.

[Bro93] J. Brown. Determination of the meter of musical scores by autocorrelation. *Journal of the Acoustical Society of America*, 94(4) :1953–1957, 1993.

[Bro99] J. Brown. Computer Identification of Musical Instruments using Pattern Recognition with Cepstral Coefficients as Features. *Journal of the Acoustical Society of America*, 105(3) :1933–1941, 1999.

[BW04] Mark A. Bartsch and Gregory H.. Wakefield. Singing voice identification using spectral envelope estimation. *IEEE Transactions Speech and Audio Processing*, 12(2) :100–109, 2004.

[Cal89] Calliope, editor. *La parole et son traitement automatique*. Masson, France, 1989.

[CG01] Wu Chou and Liang Gu. Robust Singing Detection in Speech/Music Discriminator Design. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2001.

[CKK98] Y.D. Cho, M.Y. Kim, and S.R. Kim. A Spectrally Mixed Exitation (SMX) Vocoder with Robust Parameter Determination. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1998.

[CL08] C. Cao and M. Li. Thinkit Audio Genre Classification System. In *MIREX 2008*, 2008.

[CLC05] Rui Cai, Lie Lu, and Lian-Hong Cai. Unsupervised Auditory Scene Categorization via Key Audio Effects and Information-Theoretic Co-Clustering. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1073–1076, 2005.

[Cle77] Thomas F. Cleveland. Acoustic properties of voice timbre types and their influence on voice classification. *Journal of the Acoustical Society of America*, 61(6) :1622–1629, 1977.

[CN09] M. Chemseddine and M. Noirhomme. Classification des émotions dans un espace à deux dimensions. In *XVIèmes Rencontres de la Société Francophone de Classification*, pages 133–136, 2009.

[CPR00] J. Carrive, F. Pachet, and R. Ronfard. Clavis - A Temporal Reasoning System for Classification of Audiovisual Sequences. In *Proceedings of Content-Based Multimedia Information Access Conference (RIAO)*, pages 1400–1415, 2000.

[CSY+08] W.-C. Chang, A. Su, C. Yeh, A. Roebel, and X. Rodet. Multiple-F0 tracking based on a high-order HMM model. In *Proc. of the 11th International Conference on Digital Audio Effects (DAFx-08)*, 2008.

[dCK02] A. de Cheveigné and H. Kawahara. YIN, a Fundamental Frequency Estimator for Speech and Music. *Journal of the Acoustical Society of America*, 111(4) :1917–1930, 2002.

[DGW04] S. Dixon, F. Gouyon, and G. Widmer. Towards Characterisation of Music via Rhythmic Patterns. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 509–516, 2004.

[DHS01] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Willey-Interscience, 2001.

[DOF+09] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David. Main Instrument Separation from Stereophonic Audio Signals using a Source/Filter Model. In *17th European Signal Processing Conference (EUSIPCO)*, 2009.

[Dra93] C. Drake. Reproduction of musical rhythms by children, adult musicians and adult non-musicians. *Perception and Psychophysics*, 53(1) :25–33, 1993.

[DRD08] J.-L. Durrieu, G. Richard, and B. David. Singer Melody Extraction in Polyphonic Signals using Source Separation Methods. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 169–172, 2008.

[Dre05] K. Dressler. Extraction of the Melody Pitch Contour from Polyphonic Audio. In *MIREX Melody Extraction Abstracts*, 2005.

[DW00] Peter Desain and William. L. Windsor, editors. *Rhythm Perception and Production*. Swets and Zeitlinger, 2000.

[EBD08] V. Emiya, R. Badeau, and B. David. Automatic Transcription of Piano Music Based on HMM Tracking of Jointly-Estimated Pitches. In *16th European Signal Processing Conference (EUSIPCO)*, 2008.

[EK56] A. Eronen and A. Klapuri. Musical Instrument Recognition using Cepstral Coefficients and Temporal Features. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 2000, 753–756.

[EKSP09] E. El-Khoury, C. Senac, and J. Pinquier. Improved Speaker Diarization System for Meetings. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4097–4100, 2009.

[EMP] Empirical Musicology Review.

[EPH01] Molly L. Erickson, Susan Perry, and Stephen Handel. Discrimination Functions : Can They Be Used to Classify Singing Voices ? *Journal of Voice*, 15(4) :492–502, 2001.

[ER02] Hassan Ezzaidi and Jean Rouat. Speech, Music and Songs Discrimination in the Context of Handsets Variability. In *In proceedings of ICSLP 2002*, pages 16–20, 2002.

[ER06] Slim Essid and Gaël Richard. Instrument Recognition in Polyphonic Music Based on Automatic Taxonomies. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1) :68–80, 2006.

[ERB05] Slim Essid, Gaël Richard, and David Bertrand. Instrument recognition in polyphonic music. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, page 245, 2005.

[Eve08] M. Every. Discriminating Between Pitched Sources in Music Audio. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2) :2008, 2008.

[Fab99] F. Fabbri. Browsing Music Spaces : Categories and the Musical Mind. In *Proceedings of the IASPM Conference*, 1999.

[FAP99] D.K. Fragoulis, J.N. Avaritsiotis, and C.N. Papaodysseus. Timbre Recognition of Single Notes using an ARTMAP Neural Network. In *Proc. of the 6th IEEE International Conference on Electronics, Circuits and Systems*, volume 2, pages 1099–1012, 1999.

[FGO+06] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. Okuno. Automatic Synchronization between Lyrics and Music CD Recordings based on Viterbi Alignment of Segregated Vocal Signals. In *Prco. of the 8th IEEE International Symposium on Multimedia (ISM'06)*, pages 257–264, 2006.

[FKG+06] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno. F0 Estimation Method for Singing Voice in Polyphonic Audio Signal based on Statistical Vocal Model and Viterbi Search. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 253–256, 2006.

[FU01] J. Foote and S. Uchihashi. The Beat Spectrum : a New Approach to Rhythm Analysis. In *IEEE International Conference on Multimedia and Expo (ICME '01)*, pages 881–884, 2001.

[GB08] E. Gómez and J. Bonada. Automatic Melodic Transcription of Flamenco Singing. In *Conference on Interdisciplinary Musicology (CIM08)*, 2008.

[GD05] Fabien Gouyon and Simon Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1) :34–54, 2005.

[Ger02] David B. Gerhard. Perceptual Features for a Fuzzy Speech-Song Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 4160–4163. IEEE, 2002.

[GGM+05] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and Gravier G. The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. In *Proceedings of the 9th European Conference on Speech Communication and Technology (InterSpeech 2005)*, pages 1149–1152, 2005.

[GKD+06] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An Experimental Comparision of Audio Tempo Induction Algorithms. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5) :1832–1844, 2006.

[GM94] M. Goto and Y. Muraoka. A Beat Tracking System for Acoustic Signals of Music. In *Proc. of the Second ACM International Conference on Multimedia*, pages 365–372, 1994.

[GM99] M. Goto and Y. Muraoka. Real-Time Beat Tracking for Drumless Audio Signals : Chord Change Detection for Musical Decisions. *Seech Communication*, 1999 :311–335, 1999.

[Gom06] E. Gomez. Tonal Description of Polyphonic Audio for Music Content Processing. *INFORMS Journal on Computing*, 18(3) :294–304, 2006.

[Got99] M. Goto. A Real-Time Music Scene Description System : Detecting Melody and Bass Lines in Audio Signals. In *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pages 31–40, 1999.

[Got01] Masataka Goto. An Audio-based Real-time Beat Tracking System for Music With or Without Drum-sounds. *Journal of New Music Research*, 30(2) :159–171, 2001.

[Got04] M. Goto. A Real-Time Music Scene Description System : Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-World Audio Signals. *Speech Communication*, 43(4) :311–329, 2004.

[GPT08] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis. Music Tracking in Audio Streams from Movies. In *IEEE 10th Workshop on Multimedia Signal Processing*, page 950, 2008.

[GR04] O. Gillet and G. Richard. Automatic Transcription of Drum Loops. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 269–272, 2004.

[GR05] O. Gillet and G. Richard. Drum Loops Retrieval from Spoken Queries. *Journal of Intelligent Information Systems*, 24(2) :159–177, 2005.

[GR08] O. Gillet and G. Richard. Transcription and Separation of Drum Signals From Polyphonic Music. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3) :529–540, 2008.

[Gus08] A.-M. Gustave. La loi des jingles, Jan 2008. http ://www.telerama.fr/radio/24158-la_loi_des_jingles.php.

[HDG03] P. Herrera, A. Dehamel, and F. Gouyon. Automatic Labeling of unpitched percussion sounds. In *114th Audio Engineering Society Convention*, 2003.

[Hev36] K. Hevner. Experimental Studies of the Elements of Expression in Music. *American Journal of Psychology*, 48 :246–268, 1936.

[HFC07] M. Hart, D. Fitzgerald, and M. Cranitch. Key Signature Estimation. In *Irish Signals and Systems Conference*, 2007.

[HJ07] X. Hu and Downie ; J. Exploring Mood Metadata : Relationships with Genre Artist and Usage Metadata. In *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR'07)*, 2007.

[HM03] S. Hainsworth and M. Macleod. Beat Tracking with Particle Filtering Algorithms. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 91–94, 2003.

[Hou86] P. Hougaard. A Class of Multivariate Failure Time Distributions. *Biometrika*, 73(3) :671–678, 1986.

[HSAG05] C. Harte, M. Sandler, S. Abdallah, and E. Gómez. Symbolic Tepresentation of Musical Chords : a Proposed Syntax for Text Annotations. In *Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, 2005.

[HSG06] C. Harte, M. Sandler, and M. Gasser. Detecting Harmonic Change in Musical AUdio. In *Audio Music Computer Multimedia Workshop (AMCMM'06)*, pages 21–26, 2006.

[HYG02] P. Herrera, A. Yeterian, and F. Gouyon. Automatic classification of drum sounds : a comparison of feature selection methods and classification techniques. In *International Conference on Music and Artificial Intelligence (ICMAI 2002*, pages 69–80, 2002.

[IK09] T. Inoshita and J. Katto. Key Estimation Using Circle of Fifths. In *Proc. of the 15th International Multimedia Modeling Conference (MMM'09)*, pages 287–297, 2009.

[IM98] P. Indyk and R. Motwani. Approximate Nearest Neighbors : Towards Removing the Curse of Dimensionality. In *30th Symposium on Theory of Computing*, 1998.

[IMK08] T. Izumitani, R. Mukai, and K. Kashino. A Background Music Detection Method Based on Robust Feature Extraction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13–16, 2008.

[Ins60] ANSI American National Standards Institute, 1960. http ://www.ansi.org.

[Izm05] Ö. Izmirli. Template Based Key Finding from Audio. In *International Computer Music Conference (ICMC2005)*, pages 211–214, 2005.

[JNM] Journal of New Music Research.

[KBT04] A. Kapur, M. Benning, and G. Tzanetakis. Query by Beatboxing : Music Information Retrieval for the DJ. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 170–177, 2004.

[KC05] Ian Kaminskyj and Tadeusz Czaszejko. Automatic Recognition of Isolated Monophonic Musical Instrument Sounds using kNNC. *Journal of Intelligent Information Systems*, 24(2) :199–221, 2005.

[KD06] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription.* Springer, 2006.

[KEA06] A. Klapuri, A. Eronen, and J. Astola. Analysis of the Meter of Acoustic Musical Signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1) :342–355, 2006.

[KGK+07] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno. Instrument Identification in Polyphonic Music : Feature Weighting to Minimize Influence of Sound Overlaps. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

[Kla03] A. Klapuri. Musical Meter Estimation and Music Transcription. In *Cambridge Music Processing Colloquium*, 2003.

[Kla06] A. Klapuri. Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes. In *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, 2006.

[KMAOF09] L. Koenig, C. Mailhes, R. André-Obrecht, and S. Fabre. A Continuous Voicing Parameter in the Frequency Domain. In *International Workshop on Speech and Computer (SPECOM'2009)*, 2009.

[KNL08] S.Z.K. Khine, T. L. Nwe, and H. Li. Singing Voice Detection in Pop Songs using Co-Training Algorithm. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1629–1632, 2008.

[KNS07] H. Kameoka, T. Nishimoto, and S. Sagayama. A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3) :982–994, 2007.

[Koh84] T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, 1984.

[KR03] Shandilya S. K. and Preeti Rao. Retrieving Pitch of the Singing Voice in Polyphonic Audio. In *Proc. of the National Conference on Communications (NCC)*, 2003.

[Kru90] C. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, 1990.

[KS96] B. Kröse and P. van der Smagt. *An Introduction to Neural Networks*. The University of Amsterdam, 8 edition, 1996.

[KW02] Y. E. Kim and B. Whitman. Singer Identificatio in Popular Music Recordings Using Voice Coding Features. In *Proc. of the 3th International Conference on Music Information Retrieval (ISMIR'02)*, 2002.

[LAOP07] H. Lachambre, R. André-Obrecht, and J. Pinquier. Singing Voice Characterization for Audio Indexing. In *15th European Signal Processing Conference (EUSIPCO)*, pages 1563–1540, 2007.

[LAOP09a] H. Lachambre, R. André-Obrecht, and J. Pinquier. Estimation des paramètres d'une loi de Weibull bivariée par la méthode des moments – Application à la séparation Monophonie / Polyphonie. In *16èmes Rencontres de la Société Francophone de Classification*, pages 109–112, 2009.

[LAOP09b] H. Lachambre, R. André-Obrecht, and J. Pinquier. Monophony vs Polyphony : A New Method Based on Weibull Bivariate Models. In *International Workshop on Content-Based Multimedia Indexing*, pages 68–72, 2009.

[LAOP09c] H. Lachambre, R. André-Obrecht, and J. Pinquier. Singing Voice Detection in Monophonic and Polyphonic Context. In *15th European Signal Processing Conference (EUSIPCO)*, pages 1344–1348, 2009.

[Lap00] E. Lapidaki. Stability of tempo perception in music listening. *Music Education Research*, 2(1) :25–44, 2000.

[LB90] J.C. Lu and G.K. Bhattacharyya. Some New Constructions of Bivariate Weibull Models. *Annals of Institute of Statistical Mathematics*, 42(3) :543–559, 1990.

[LCB99] A. Loscos, P. Cano, and J. Bonada. Low-Delay Singing Voice Alignment to Text. In *International Computer Music Conference (ICMC2005)*, 1999.

[LD92] R. Larsen and E. Diener. Problems and Promises with the Circumplex Model of Emotion. *Review of Personality and Social Psychology*, 13 :25–59, 1992.

[LDB07] J. Louradour, K. Daoudi, and F. Bach. Feature Space Mahalanobis Sequence Kernels : Application to SVM Speaker Verification. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8) :2465–2475, 2007.

[LE08] K. Lee and D. Ellis. Detecting Music in Ambient Audio by Long-Window Autocorrelation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9–12, 2008.

[LGD07] H. Lukashevich, M. Gruhne, and C. Dittmar. Effective Singing Voice Detection in Popular Music using ARMA Filtering. In *Proc. of the 10th International Conference on Digital Audio Effects (DAFx-07)*, 2007.

[LKS+98] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney. Classification of Audio Signals Using Statistical Features on Time and Wavelet Transform Domains. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3621–3624, 1998.

[LLZ06] L. Lu, D. Liu, and H.-J. Zhang. Automatic Mood Detection and Tracking of Music Audio Signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1) :5–18, 2006.

[LO03] T. Li and M. Ogihara. Detecting Emotion in Music. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 239–240, 2003.

[LOL03] T. Li, M. Ogihara, and Q. Li. A Comparative Study on Content-Based Music Genre Classification. In *SIGIR'03*, pages 282–289, 2003.

[LR04] Arie Livshin and Xavier Rodet. Musical Instrument Identification in Continuous Recordings. In *Proc. of the 7th International Conference on Digital Audio Effects (DAFx-04)*, 2004.

[LR05] T. Lidy and A. Rauber. Evaluation of Feature Extractors and Psych-Acoustic Transformations for Music Genre Classification. In *Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, 2005.

[LR06] Arie Livshin and Xavier Rodet. The Significance of the Non-Harmonic "Noise" Versus the Harmonic Series for Musical Instrument Recognition. In *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, 2006.

[LRP+08] T. Lidy, A. Rauber, A. Pertusa, P. Ponce de León, and J. Iñesta. Audio Music Classification using a Combination of Spectral, Timbral, Rhythmic, Temporal and Symbolic Features. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, 2008.

[LS06] K. Lee and M. Slaney. Automatic Chord Recognition from Audio Using a Supervised HMM Trained with Audio-from-Symbolic Data. In *Audio Music Computer Multimedia Workshop (AMCMM'06)*, pages 11–20, 2006.

[LS08] K. Lee and M. Slaney. Acoustic Chord Transcription and Key Extraction From Audio Using Key-Dependant HMMs Trained on Synthesized Audio. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2) :291–301, 2008.

[LSC+09] T. Lidy, C. Silla, O. Cornelis, F. Gouyon, A. Rauber, C. Kaestner, and A. Koerich. On the Suitability of State-of-the-Art Music Information Retrival Methods for Analysing, Categorizing and Accessing Non-Western and Ethnic Music Collections. *Signal Processing*, 2009.

[Mar04a] M. Marolt. A Connectionist Approach to Automatic Transcription of Polyphonic Piano Music. *IEEE Transaction on Multimedia*, 6(3) :439–449, 2004.

[Mar04b] M. Marolt. On Finding Melodic Lines in Audio Recordings. In *Proc. of the 7th International Conference on Digital Audio Effects (DAFx'04)*, pages 217–221, 2004.

[Mar04c] Jeremy Marozeau. *L'effet de la fréquence fondamentale sur le timbre*. PhD thesis, Université Pierre et Marie Curie, Paris VI, 2004.

[MBTL07] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange. Polyphonic Instrument Recognition using Spectral Clustering. In *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR'07)*, 2007.

[MDH+07] M. Mauch, S. Dixon, C. Harte, M. Casey, and B. Fields. Discovering Chord Idioms Through Beatles and Real Book Songs. In *Proc. of the 8th International Conference on Music Information Retrieval (ISMIR'07)*, 2007.

[ME08] M. M.Mandel and D. Ellis. Labrosas's Audio Classification Submissions. In *MIREX 2008*, 2008.

[MF06] C. McKay and I. Fujinaga. Musical Genre Classification : Is it Worth Pursuing and How Can it Be Improved ? In *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, pages 101–106, 2006.

[MH00] Y. Meron and K. Hirose. Synthesis of Vibrato Singing. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 745–748, 2000.

[MHS08] M. Markaki, A. Holzapfel, and Y. Stylianou. Singing Voice Detection using Modulation Frequency Features. In *Workshop on Statistical And Perceptual Audition (SAPA)*, 2008.

[MM06] A. Mesaros and S. Moldovan. Method for Singing Voice Identification using Energy Coefficients as Features. In *IEEE International Conference on Automation, Quality and Testing, Robotics*, pages 161–166, 2006.

[Moo77] J. Moorer. On the Transcription of Musical Sound bu Computer. *Computer Music Journal*, 1(4) :32–38, 1977.

[Mor68] E. Morice. Quelques problèmes d'estimation relatifs à la loi de Weibull. *Revue de statistique appliquée*, 16(3) :43–63, 1968.

[MP] Music Perception.

[MS] Musicae Scientiae.

[MV08] A. Mesaros and T. Virtanen. Automatic Alignment of Music Audio and Lyrics. In *Proc. of the 11th International Conference on Digital Audio Effects (DAFx'08)*, 2008.

[MV09] A. Messaros and T. Virtanen. Adaptation of a Speech Recognizer for Singing Voice. In *17th European Signal Processing Conference (EUSIPCO)*, pages 1779–1783, 2009.

[MWXW04] N.C. Maddage, Kongwah Wan, Changsheng Xu, and Ye Wang. Singing Voice Detection using Twice-Iterated Composite Fourier Transform. In *IEEE International Conference on Multimedia and Expo (ICME '04)*, volume 2, pages 1347–1350, 2004.

[MXW04] N. C. Maddage, C. Xu, and Y. Wang. Singer Identification Based on Vocal and Instrumental Models. In *Proc. of the 17th International Conference on Pattern Recognition (ICPR'04)*, pages 375–378, 2004.

[NAW01] S. Nawab, S. Ayyash, and R. Wotiz. Identification of Musical Chords Using Constant-Q Spectra. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3373–3376, 2001.

[NL07a] T. Nwe and H. Li. Singing Voice Detection using Perceptually-Motivated Features. In *Proceedings of the 15th international conference on Multimedia (MM'07)*, pages 309–312, 2007.

[NL07b] T. L. Nwe and H. Li. Exploring Vibrato-Motivated Acoustic Features for Singer Identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2) :519–530, 2007.

[NS06] K. Noland and M. Sandler. Key Estimation Using a Hidden Markov Model. In *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, pages 121–126, 2006.

[NSW04] T. L. Nwe, A. Shenoy, and Y. Wang. Singing Voice Detection in Popular Music. In *Proceedings of the 12th annual ACM International Conference on Multimedia (Multimedia'04)*, pages 324–327, 2004.

[OGIT05] Y. Ohishi, M. Goto, K. Itou, and K. Tekada. Discrimination between Singing and Speaking Voices. In *Interspeech - European Conference on Speech Communication and Technology*, 2005.

[PAO04] J. Pinquier and R. André-Obrecht. Jingle Detection and Identification in Audio Documents. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume IV, pages 329–332. IEEE, 2004.

[Pau04] S. Pauws. Musical Key Extraction from Audio. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, 2004.

[PC00] F. Pachet and D. Cazaly. A Taxonomy of Musical Genre. In *Proc. of Content-Based Multimedia Information Access (RIAO)*, 2000.

[PdLIn07] P. Ponce de León and J. Iñesta. Pattern Recognition Approach for Music Style Identification Using Shallow Statistical Descriptors. *IEEE Transactions on Systems, Man, and Cybernetics – Part C : Applications and Reviews*, 37(2) :248–157, 2007.

[PE05] G. Poliner and D. Ellis. A Classification Approach to Melody Transcription. In *Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, pages 161–166, 2005.

[PE07] G. Poliner and P. Ellis. A Discriminative Model for Polyponic Piano Transcription. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

[Pee05a] G. Peeters. Rhythm Classification using Spectral Rhythm Patterns. In *Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, 2005.

[Pee05b] G. Peeters. Time variable tempo detection and beat marking. In *International Computer Music Conference (ICMC'05)*, 2005.

[Pee06] G. Peeters. Chroma-Based Estimation of Musical Key from Audio-Signal Analysis. In *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, 2006.

[Pee07a] Geoffroy Peeters. Template-Based Estimation of Time-Varying Tempo. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

[PEE+07b] G. Poliner, D. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody Transcription From Music Audio : Approaches and Evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4) :1247–1256, 2007.

[Pee08] G. Peeters. A Generic Training and Classification System for MIREX08 Classification Tasks : Audio Music Mood, Audio Genre, Audio Artist and Audio Tag. In *MIREX 2008*, 2008.

[PK03] J. Paulus and A. Klapuri. Model-Based Event Labeling in the Transcription of Percussive Audio Signals. In *Proc. of the 6th International Conference on Digital Audio Effects (DAFx-03)*, pages 73–77, 2003.

[PMC04] R. Paiva, T. Mendes, and A. Cardoso. A Methodology for Detection of Melody in Polyphonic Musical Signals. In *116th Audio Engineering Society Convention*, 2004.

[PoM] Psychology of Music.

[PP07] H. Papadopoulos and G. Peeters. Large-Scale Study of Chord Estimation Agorithmes Based on Chroma Representation and HMM. In *International Workshop on Content-Based Multimedia Indexing*, pages 53–60, 2007.

[PP08] H. Papadopoulos and G. Peeters. Simultaneous Estimation of Chord Progression and Downbeats from an Audio File. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–124, 2008.

[PR02] G. Peeters and X. Rodet. Automatically selecting signal descriptors for sound classification. In *International Computer Music Conference (ICMC2002)*, pages 455–458, 2002.

[PRAO03] J. Pinquier, J.L. Rouas, and R. André-Obrecht. A Fusion Study in Speech / Music Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 17–20. IEEE, 2003.

[Rap02] C. Raphael. Automatic Transcription of Piano Music. In *Proc. of the 3th International Conference on Music Information Retrieval (ISMIR'02)*, 2002.

[RDS+99] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette. Vibraot : Detection, Estimation, Extraction, Modification. In *Proc. of the 2nd International Conference on Digital Audio Effects (DAFx'99)*, 1999.

[RF01] A. Rauber and M. Frühwirth. Automatically Analyzing and Organizing Music Archives. In *Proc. of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*, 2001.

[RFPAO05] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht. Rhythmic Unit Extraction and Modelling for Automatic Language Identification. *Speech Communication*, 47(4) :436–456, 2005.

[RH07] M. Rocamora and P. Herrera. Comparing Audio Descriptors for Singing Voice Detection in Music Audio Files. In *Brazilian Symposium on Computer Music, 11th. San Pablo, Brazil*, 2007.

[RK05] M. Ryynänen and A. Klapuri. Polyphonic Music Transcription using Note Event Modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 319–322, 2005.

[RK06] M. Ryynänen and A. Klapuri. Transcription of the Singing Melody in Polyphonic Music. In *Proc. of the 7th International Conference on Music Information Retrieval (ISMIR'06)*, pages 222–227, 2006.

[RP09] L. Regnier and G. Peeters. Singing Voice Detection in Music Tracks using Direct Voice Vibrato Detection. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1685–1688, 2009.

[RR05] A. Röebel and X. Rodet. Efficient Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope Preservation. In *Proc. of the 8th International Conference on Digital Audio Effects (DAFx'05)*, 2005.

[RRD08] M. Ramona, G. Richard, and B. David. Vocal Detection in Music with Support Vector Machines. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1885–1888, 2008.

[Rus80] J. Russel. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39 :1161–1178, 1980.

[SC00] Hari Sundaram and Shih-Fu Chang. Audio Scene Segmentation using Multiple Features, Models and Time Scales. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2441–2444, 2000.

[Sch97] E. Scheirer. Pulse Tracking with a Pitch Tracker. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[Sch99] E. Schubert. *Measurement and Time Series Analysis of Emotion in Music*. PhD thesis, University of New South Wales, 1999.

[SE03] A. Sheh and D. Ellis. Chord Segmentation and Recognition using EM-Trained Hidden Markov Models. In *Proc. of the 4th International Conference on Music Information Retrieval (ISMIR'03)*, 2003.

[SE07] C. Smit and D. Ellis. Solo Voice Detection via Optimal Cancelation. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007.

[Sea38] C. E. Seashore. *Psychology of Music*. McGraw-Hill Book Company, inc., 1938.

[She64] R. Shepard. Circularity in Judgments of Relative Pitch. *Journal of the Acoustical Society of America*, 1964.

[SJ01] B. Su and S.-K. Jeng. Multi-Timbre Chord Classification Using Wavelet Transform and Self-Organized Map Neural Network. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3377–3380, 2001.

[SOK06] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *MIR '06 : Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, 2006.

[SPS07] K. Seyerlehner, T. Pohle, and M. Schedl. Automatic Music Detection in Television Productions. In *Proc. of the 10th International Conference on Digital Audio Effects (DAFx'07)*, 2007.

[SRRR09] S. Santosh, S. Ramakrishnan, Vishweshwara Rao, and Preeti Rao. Improving Singing Voice Detection in Presence of Pitched Accompaniment. In *Proc. of the National Conference on Communications (NCC)*, 2009.

[SS97] E. Scheirer and M. Slaney. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1997.

[SSWW98] H. Soltau, T. Schultz, M. Westphal, and A. Waibel. Recognition of Music Types. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1137–1140, 1998.

[Sun94] J. Sundberg. Acoustic and Psychoacoustic Aspects of Vocal Vibrato. *STL-QPSR*, 35(2–3) :45–68, 1994.

[SW05] A. Shenoy and Y. Wang. Key, Chord and Rhythm Tracking of Popular Music Recordings. *Computer Music Journal*, 29(3) :75–86, 2005.

[TAO+05] Toru Taniguchi, Akishige Adachi, Shigeki Okawa, Masaaki Honda, and Katsuhiko Shirai. Discrimination of Speech, Musical Istruments and Singing Voices Using the Temporal Patterns of Sinusoidal Segments in Audio Signals. In *Interspeech - European Conference on Speech Communication and Technology*. ISCA, September 2005.

[TD00] R. Timmers and P. Desain. Vibrato : Questions and Answers from Musicians and Science. In *Proc. Int. Conf. on Music Perception and Cognition*, 2000.

[TDDB05] K. Tanghe, S. Degroeve, and B. De Baets. An Algorithm for Detecting and Labeling Drum Events in Polyphonic Music. In *Proc. of 2005 MIREX Evaluation Campaign*, 2005.

[Tem01] D. Temperley. *The Cognition of Basic Musical Structures*. The MIT Press, 2001.

[TLL08] W.-H. Tsai, S.-J. Liao, and C. Lai. Automatic Identification of Simultaneous Singer Recordings. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, pages 115–120, 2008.

[TTK05] M. Tolos, R. Tato, and T. Kemp. Mood-Based Navigation Trough Large Collections of Musical Data. In *Consumer Communication and Network Conference (CCNC 2005)*, pages 71–75, 2005.

[TTKV08] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multi-Label Classification of Music into Emotions. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, pages 325–330, 2008.

[Tza04] G. Tzanetakis. Musical Genre Classification of Audio Signals. *IEEE Transactions on Audio, Speech and Language Processing*, 10(5) :293–302, 2004.

[Tza08] G. Tzanetakis. Marsyas Submissions to MIREX 2007. In *MIREX 2008*, 2008.

[TZW08] Chee Chuan Toh, Bingjun Zhang, and Ye Wang. Multiple-Feature Fusion Based Onset Detection for Solo Singing Voice. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, pages 515–520, 2008.

[VAB09] E. Vincent, S. Araki, and P. Bofill. The 2008 Signal Separation Evaluation Campaign : A Community-Based Approach to Large Scale Evaluation. In *International Conference on Independant Component Analysis (ICA2009)*, 2009.

[Vap00] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2nde edition, 2000.

[Var04] Various. *The Real Book*. Hal Leonard Corporation, 6th edition, 2004.

[VGD05] V. Verfaille, C. Guastavino, and P. Depalle. Perceptual Evaluation of Vibrato Models. In *Proceedings of the Conference on Interdisciplinary Musicology*, pages 149–151, 2005.

[Vib09] Flute Vibrato, 2009. www.standingstones.com/flutevib.html.

[VP05] E. Vincent and M. Plumbley. Predominant-F0 Estimation using Bayesian Harmonic Waveform Models. In *MIREX Melody Extraction Abstracts*, 2005.

[VR05] E. Vincent and X. Rodet. Instrument Identification in Solo and Ensemble Music using Independent Subspace Analysis. In *Proc. of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, pages 576–581, 2005.

[WB03] Gregory H. Wakefield and A. Bartsch, Mark. Where's Caruso ? Singer Identification by Listener and Machine. In *Cambridge Music Processing Colloquium*, 2003.

[Wik] http ://en.wikipedia.org.

[WLC03] C. Wang, R. Lyu, and Y. Chiang. An Automatic Singing Transcription System with Multilingual Singing Lyrics Recognizer and Robust Melody Tracker. In *EUROSPEECH 2003*, 2003.

[Yeh08] Chunghsin Yeh. *Multiple Fundamental Frequency Estimation of Polyphonic Recordings*. PhD thesis, Université Paris VI - Pierre et Marie Curie, 2008.

[YGO04] K. Yoshii, M. Goto, and H. G. Okuno. Automatic Drum Sound Description for Real-World Music using Template Adaptation and Matching Methods. In *Proc. of the 5th International Conference on Music Information Retrieval (ISMIR'04)*, pages 184–191, 2004.

[YLSC08] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen. A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2) :448–457, 2008.

[ZG08] X. Zhang and D. Gerhard. Chord Recognition using Instrument Voicing Constraints. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR'08)*, pages 33–38, 2008.

[Zha03] Tong Zhang. Automatic Singer Identification. In *IEEE International Conference on Multimedia and Expo (ICME '03)*, pages 33–36, 2003.

[ZPDG02] A. Zils, F. Pachet, O. Delerue, and F. Gouyon. Automatic Extraction of Drum Tracks from Polyphonic Music Signals. In *International Conference on Web Delivering of Music (WEDELMUSIC2002)*, pages 179–183, 2002.

Résumé

Actuellement, la quantité de musique disponible, notamment via Internet, va tous les jours croissant. Les collections sont trop gigantesques pour qu'il soit possible d'y naviguer ou d'y rechercher un extrait sans l'aide d'outils informatiques. Notre travail se place dans le cadre général de l'indexation automatique de la musique.

Afin de situer le contexte de travail, nous proposons tout d'abord une brève revue des travaux réalisés actuellement pour la description automatique de la musique à des fins d'indexation : reconnaissance d'instruments, détermination de la tonalité, du tempo, classification en genre et en émotion, identification du chanteur, transcriptions de la mélodie, de la partition, de la suite d'accords et des paroles. Pour chacun de ces sujets, nous nous attachons à définir le problème, les termes techniques propres au domaine, et nous nous attardons plus particulièrement sur les problèmes les plus saillants.

Dans une seconde partie, nous décrivons le premier outil que nous avons développé : une distinction automatique entre les sons monophoniques et les sons polyphoniques. Nous avons proposé deux nouveaux paramètres, basés sur l'analyse d'un indice de confiance. La modélisation de la répartition bivariée de ces paramètres est réalisée par des distributions de Weibull bivariées. Le problème de l'estimation des paramètres de cette distribution nous a conduit à proposer une méthode originale d'estimation dérivée de l'analyse des moments de la loi. Une série d'expériences nous permet de comparer notre système à des approches classiques, et de valider toutes les étapes de notre méthode.

Dans la troisième partie, nous proposons une méthode de détection du chant, accompagné ou non. Cette méthode se base sur la détection du vibrato, un paramètre défini à partir de l'analyse de la fréquence fondamentale, et défini *a priori* pour les sons monophoniques. À l'aide de deux segmentations, nous étendons ce concept aux sons polyphoniques, en introduisant un nouveau paramètre : le vibrato étendu. Les performances de cette méthode sont comparables à celles de l'état de l'art. La prise en compte du pré-traitement monophonique / polyphonique nous a amenés à adapter notre méthode de détection du chant à chacun de ces contextes. Les résultats s'en trouvent améliorés.

Après une réflexion sur l'utilisation de la musique pour la description, l'annotation et l'indexation automatique des documents audiovisuels, nous nous posons la question de l'apport de chacun des outils décrits dans cette thèse au problème de l'indexation de la musique, et de l'indexation des documents audiovisuels par la musique et offrons quelques perspectives.

Mots-clés: Musique, Indexation, Monophonie, Polyphonie, Chant, Loi de Weibull bivariée, Vibrato.

Abstract

Currently, the amount of music available, notably via the Internet, is growing daily. The collections are too huge for a user to navigate into without help from a computer. Our work takes place in the general context of music indexation.

In order to detail the context of our work, we present a brief overview of the work currently made in music indexation for indexation : instrument recognition, tonality and tempo estimation, genre and mood classification, singer identification, melody, score, chord and lyrics transcription. For each of these subjects, we insist on the definition of the problem and of technical terms, and on the more imporants problems encountered.

In a second part, we present au method we developped to automatically distinguish between monophonic and polyphonic sounds. For this task, we developped two new parameters, based on the analysis of a confidence indicator. The modeling of these parameters is made with Weibull bivariate distributions. We studied the problem of the estimation of the parameters of this distribution, and suggested an original method derived from the moment method. A full set of experiment allow us to compare our system with classical method, and to validate each step of our approach.

In the third part, we present a singing voice detector, in monophonic and polyphonic context. This method is base on the detection of vibrato. This parameter is derived from the analysis of the fundamental frequency, so it is *a priori* defined for monophonic sounds. Using two segmentations, we extend this concept to polyphonic sounds, and present a new parameter : the extended vibrato. Our system's performances are comparable with those of state-of-the-art methods. Using the monophonic / polyphonic distinction as a pre-processing allow us to adapt our singing voice detector to each context. This leads to an improvment of the results.

After giving some reflexions on the use of music for automatic description, annotating and indexing of audiovisual documents, we present the contribution of each tool we presented to music indexation, and to audiovisual documents indexation using music, and finally give some perspectives.

Keywords: Music, Indexation, Monophony, Polyphony, Singing voice, Weibull bivariate distribution, Vibrato.